# CS 1674: Vision Language Models

**PhD. Nils Murrugarra-Llerena**
nem177@pitt.edu



# Plan for this lecture

1. What is Natural Language Generation (NLG)?

2. Multimodal Large Language Models (MLLM)

3. Functionality
   - Contrastive Learning Models
   - Generative Models

# What is Natural Language Generation (NLG)?

Natural language generation is one side of natural language processing. NLP =

Natural Language Understanding (NLU) +
Natural Language Generation (NLG)



NLG focuses on systems that produce fluent, coherent and useful language output for human consumption

Deep Learning is powering next-gen NLG systems!

Xiang Lisa Li, Antoine Bosselut, Chris Maning - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# Examples of NLG

Machine Translation systems

input: utterances in source languages

output: translated text in target languages.





Digital assistant (dialogue) systems use NLG

input: dialog history

output: text that respond / continue the conversation



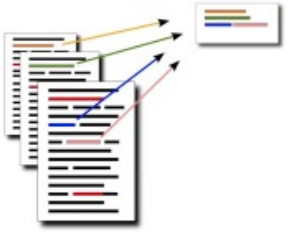


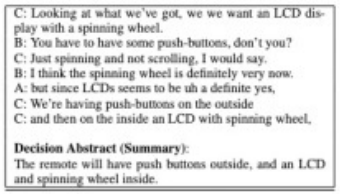



Summarization systems (for research articles, email, meetings, documents) use NLG

input: long documents

output: summarization of the long documents

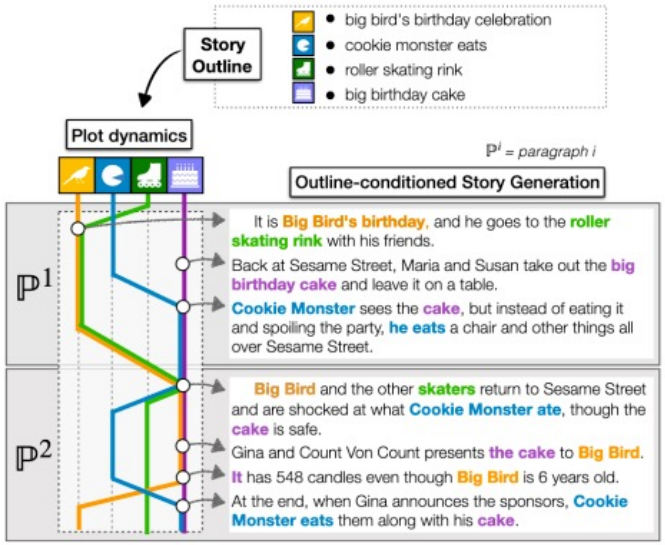





Xiang Lisa Li, Antoine Bosselut, Chris Maning - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# More Interesting Examples of NLG

## Creative stories




(Rashkin et al.., EMNLP 2020)

## Data-to-text

**Table Title:** Robert Craig (American football)
**Section Title:** National Football League statistics
**Table Description:** None

| | | RUSHING | | | | | RECEIVING | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| YEAR | TEAM | ATT | YDS | AVG | LNG | TD | NO. | YDS | AVG | LNG | TD |
| 1983 | SF | 176 | 725 | 4.1 | 71 | 8 | 48 | 427 | 8.9 | 23 | 4 |
| 1984 | SF | 155 | 649 | 4.2 | 28 | 4 | 71 | 675 | 9.5 | 64 | 3 |
| 1985 | SF | 214 | 1050 | 4.9 | 62 | 9 | 92 | 1016 | 11 | 73 | 6 |
| 1986 | SF | 204 | 830 | 4.1 | 25 | 7 | 81 | 624 | 7.7 | 48 | 0 |
| 1987 | SF | 215 | 815 | 3.8 | 25 | 3 | 66 | 492 | 7.5 | 35 | 1 |
| 1988 | SF | 310 | 1502 | 4.8 | 46 | 9 | 76 | 534 | 7.0 | 22 | 1 |
| 1989 | SF | 271 | 1054 | 3.9 | 27 | 6 | 49 | 473 | 9.7 | 44 | 1 |
| 1990 | SF | 141 | 439 | 3.1 | 26 | 1 | 25 | 201 | 8.0 | 31 | 0 |
| 1991 | RAI | 162 | 590 | 3.6 | 15 | 1 | 17 | 136 | 8.0 | 20 | 0 |
| 1992 | MIN | 105 | 416 | 4.0 | 21 | 4 | 22 | 164 | 7.5 | 22 | 0 |
| 1993 | MIN | 38 | 119 | 3.1 | 11 | 1 | 19 | 169 | 8.9 | 31 | 1 |
| Totals | - | 1991 | 8189 | 4.1 | 71 | 56 | 566 | 4911 | 8.7 | 73 | 17 |

Craig finished his eleven NFL seasons with 8,189 rushing yards and 566 receptions for 4,911 receiving yards.

(Parikh et al.., EMNLP 2020)

## Visual description



Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

(Krause et al. CVPR 2017)

Xiang Lisa Li, Antoine Bosselut, Chris Maning - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# SOTA NLG System

ChatGPT is an NLG system!

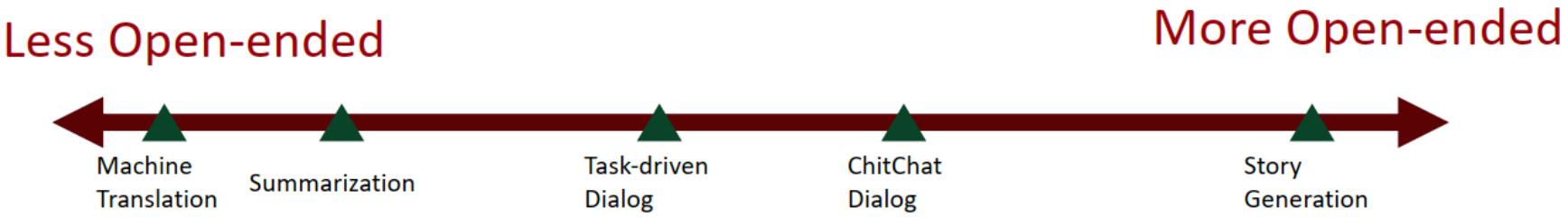
It's general purpose and can do many NLG tasks!

XL Got any creative ideas for a 10 year old's birthday?

Here are some creative birthday ideas for a 10 year old:

1. Scavenger Hunt: A birthday scavenger hunt can be a fun and interactive activity for kids. You can create a series of clues that lead the birthday child to the next one, until they reach the final prize.
2. DIY Crafts: Crafts are always a hit with kids. You can arrange a crafting session where the birthday child and their friends can make their own personalized gifts or decorations.
3. Movie Night: Have a movie night with friends and family at home or at a local theater. You can also make it a sleepover by setting up sleeping bags, pillows, and blankets.

# NLG Systems




Open-ended generation: the output distribution still has high freedom

Non-open-ended generation: the input mostly determines the output generation.

Xiang Lisa Li, Antoine Bosselut, Chris Maning - https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# Plan for this lecture

1. What is Natural Language Generation (NLG)?

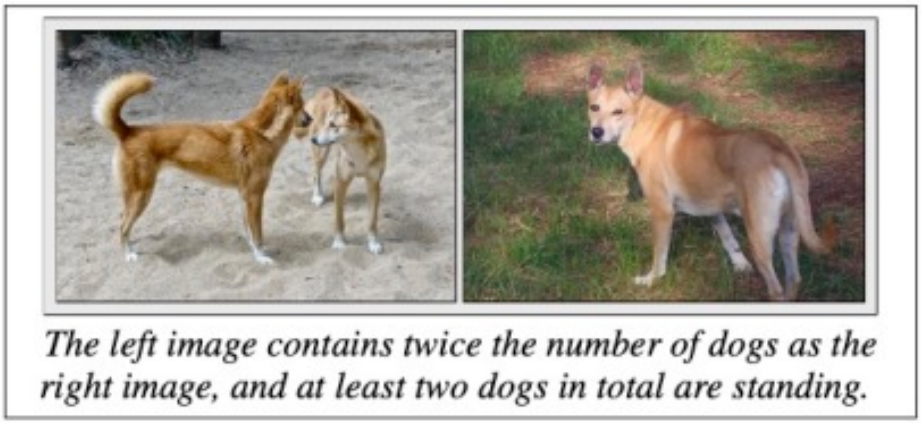
2. Multimodal Large Language Models (MLLM)

3. Functionality
    - Contrastive Learning Models
    - Generative Models

# Vision-Language Models Tasks

## Visual reasoning

Given an image (or a pair of images) determine if some natural language statement about the image(s) is true or false



*The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.*



*One image shows exactly two brown acorns in back-to-back caps on green foliage.*

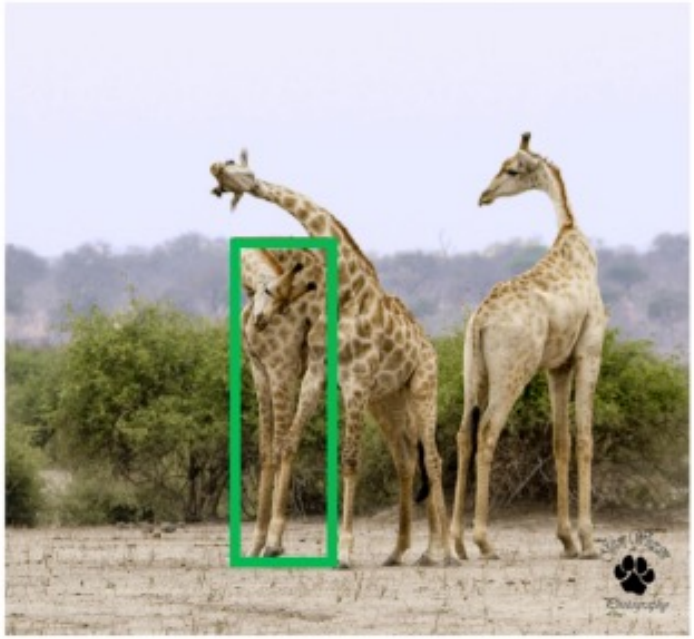
https://www.cs.cmu.edu/~mgormley/courses/10423/

Source: https://aclanthology.org/P19-1644.pdf

# Vision-Language Models Tasks

## Visual grounding

Locate an object in some image given a natural language description



RefCOCO:
1. giraffe on left
2. first giraffe on left

RefCOCO+:
1. giraffe with lowered head
2. giraffe head down

RefCOCOg:
1. an adult giraffe scratching its back with its horn
2. giraffe hugging another giraffe

https://www.cs.cmu.edu/~mgormley/courses/10423/

Source: https://aclanthology.org/P19-1644.pdf

# Vision-Language Models Tasks

## Visual question answering

Given an image (or images), respond to arbitrary, potentially open-ended



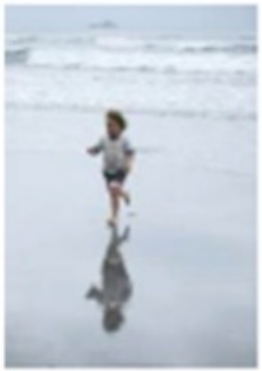

https://www.cs.cmu.edu/~mgormley/courses/10423/

Source: https://aclanthology.org/P19-1644.pdf

# Vision-Language Models Tasks

## Caption generation

Create natural language descriptions of content of some image



**Ground Truth Caption:** A little boy runs away from the approaching waves of the ocean.

**Generated Caption:** A young boy is running on the beach.



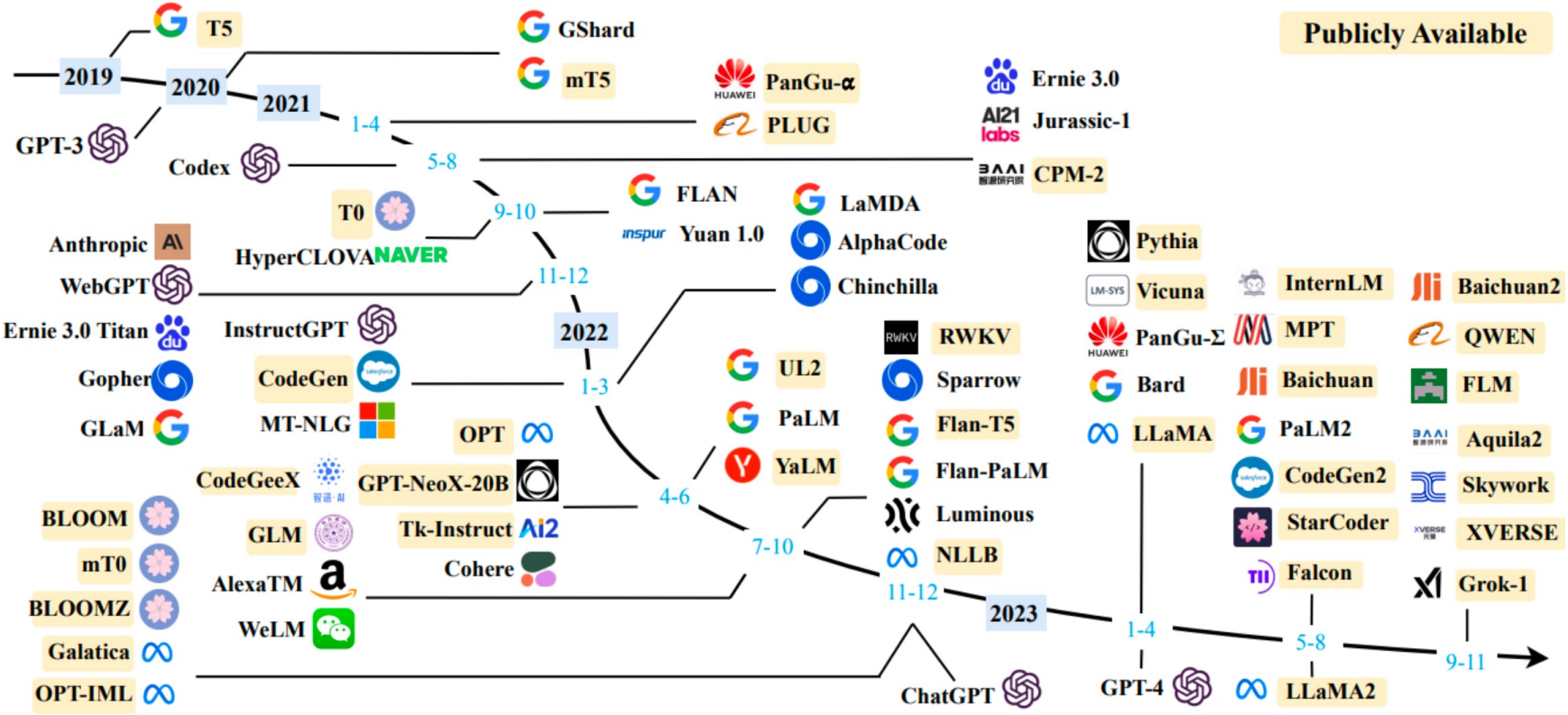
**Ground Truth Caption:** A brunette girl wearing sunglasses and a yellow shirt.

**Generated Caption:** A woman in a black shirt and sunglasses smiles.

https://www.cs.cmu.edu/~mgormley/courses/10423/

Source: https://aclanthology.org/P19-1644.pdf

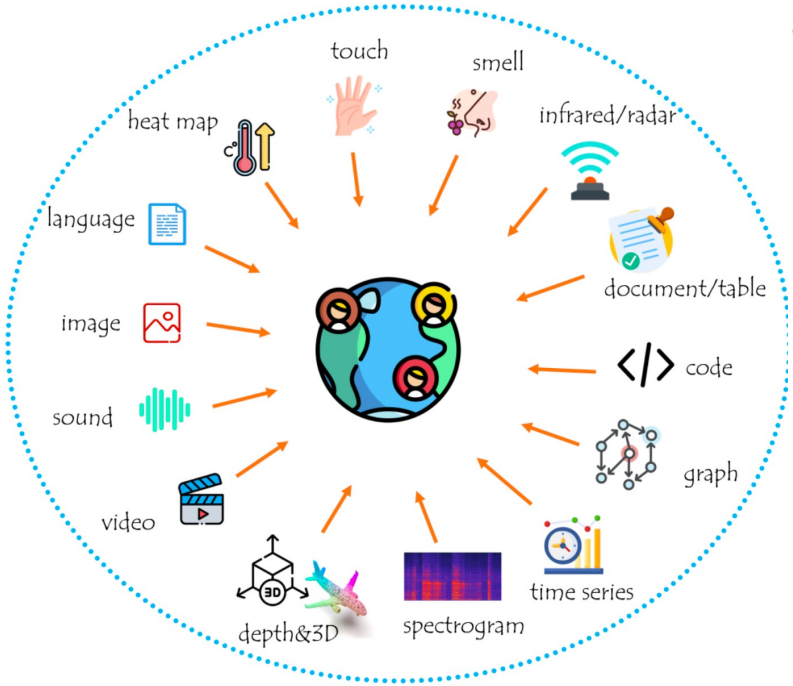
# Very Rapid Evolvement of Language-based LLMs




[1] A Survey of Large Language Models. https://github.com/RUCAIBox/LLMSurvey , 2023

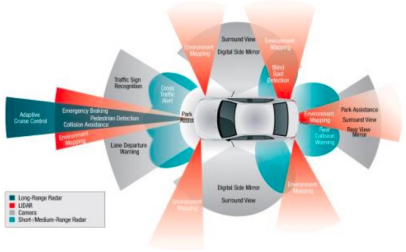
# Expanding to Multimodality

This world we live in is replete with multimodal information & signals, not just language.



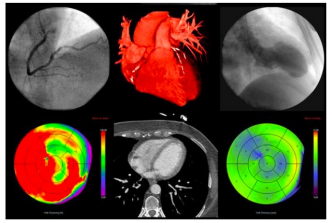

- **Autonomous Driving Systems**

  *In this application, vehicles use a combination of visual data (cameras), spatial data (LiDAR), and auditory signals (sonar) to navigate safely.*



- **Healthcare Diagnostics**

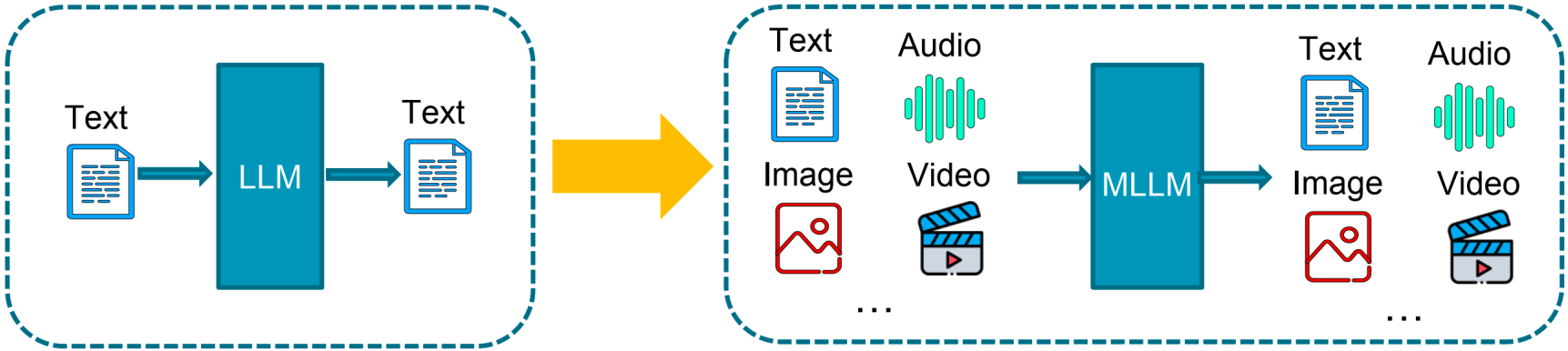
  *Medical **imaging** tools like MRI, CT scans, and X-rays, along with patient history and verbal symptoms, are used to diagnose diseases.*



[1] A Survey of Large Language Models. https://github.com/RUCAIBox/LLMSurvey , 2023

# Building Multimodal LLMs

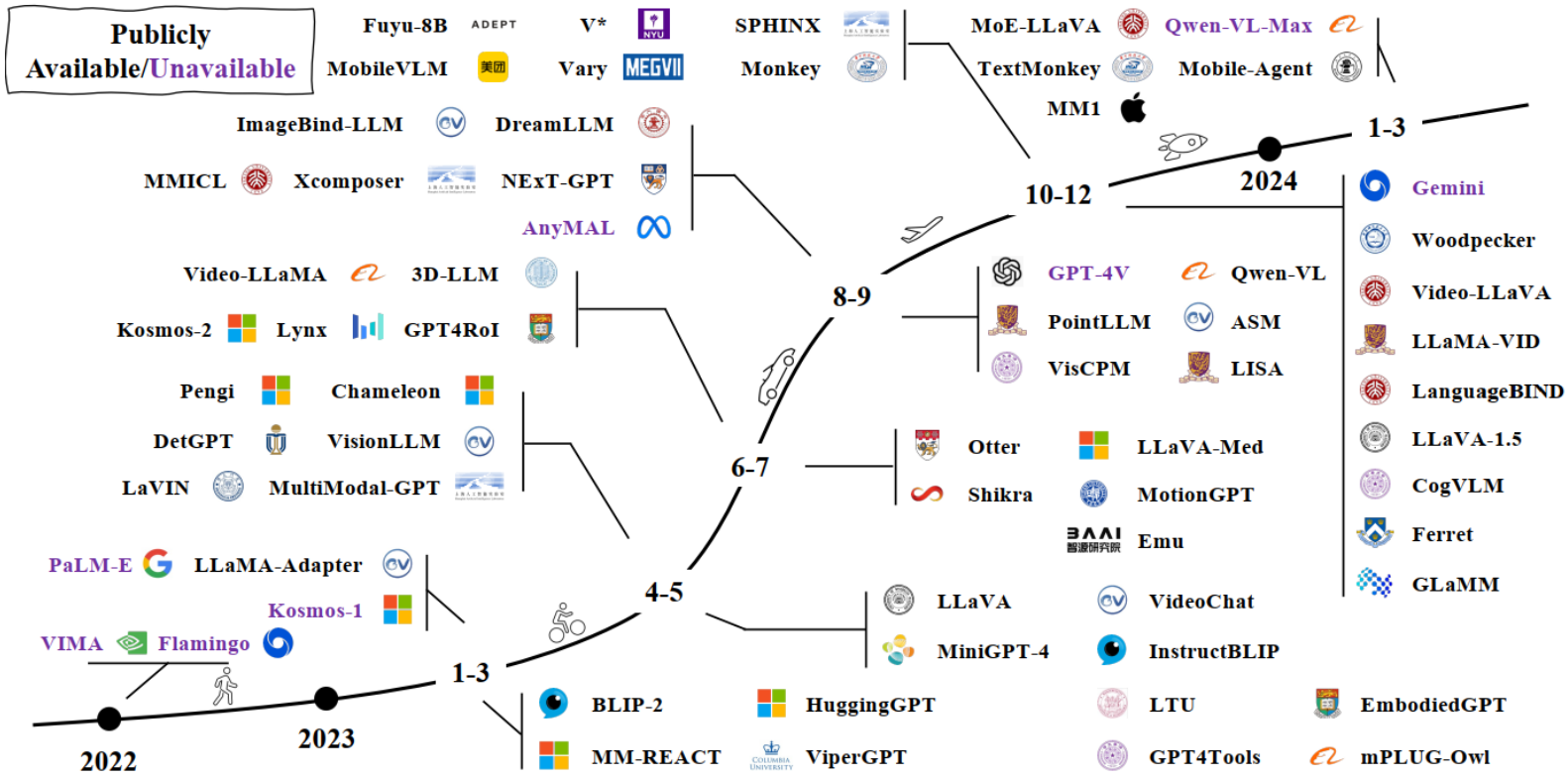
Can we transfer the success of LLMs to MLLMs, enabling LLMs to comprehend multimodal information as deeply as they understand language?




Perceiving and interacting with the world as HUMAN BEINGs do, might be the key to achieving human-level AI

Multimodal Large Language Models (MLLM) – Tutorial - https://mllm2024.github.io/CVPR2024/

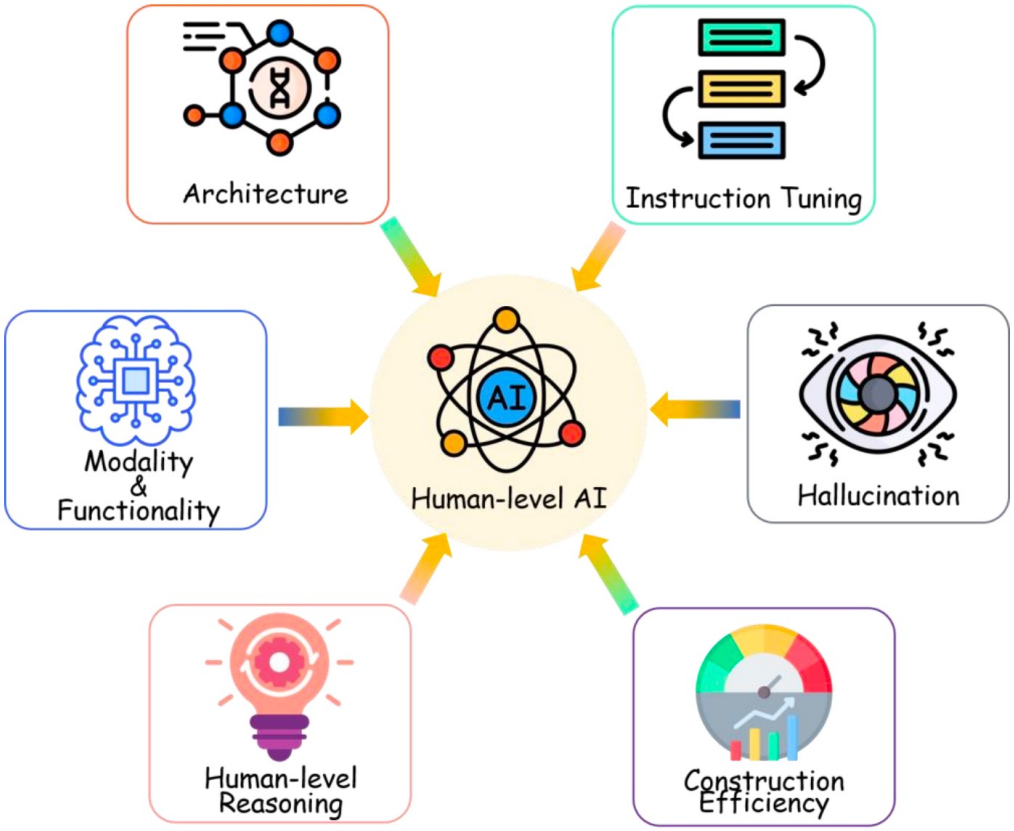
# Building Multimodal LLMs




[1] A Survey on Multimodal Large Language Models. https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models , 2023.

Multimodal Large Language Models (MLLM) – Tutorial - https://mllm2024.github.io/CVPR2024/

# Key Aspects for Building Powerful MLLMs




Multimodal Large Language Models (MLLM) – Tutorial - https://mllm2024.github.io/CVPR2024/

# Plan for this lecture

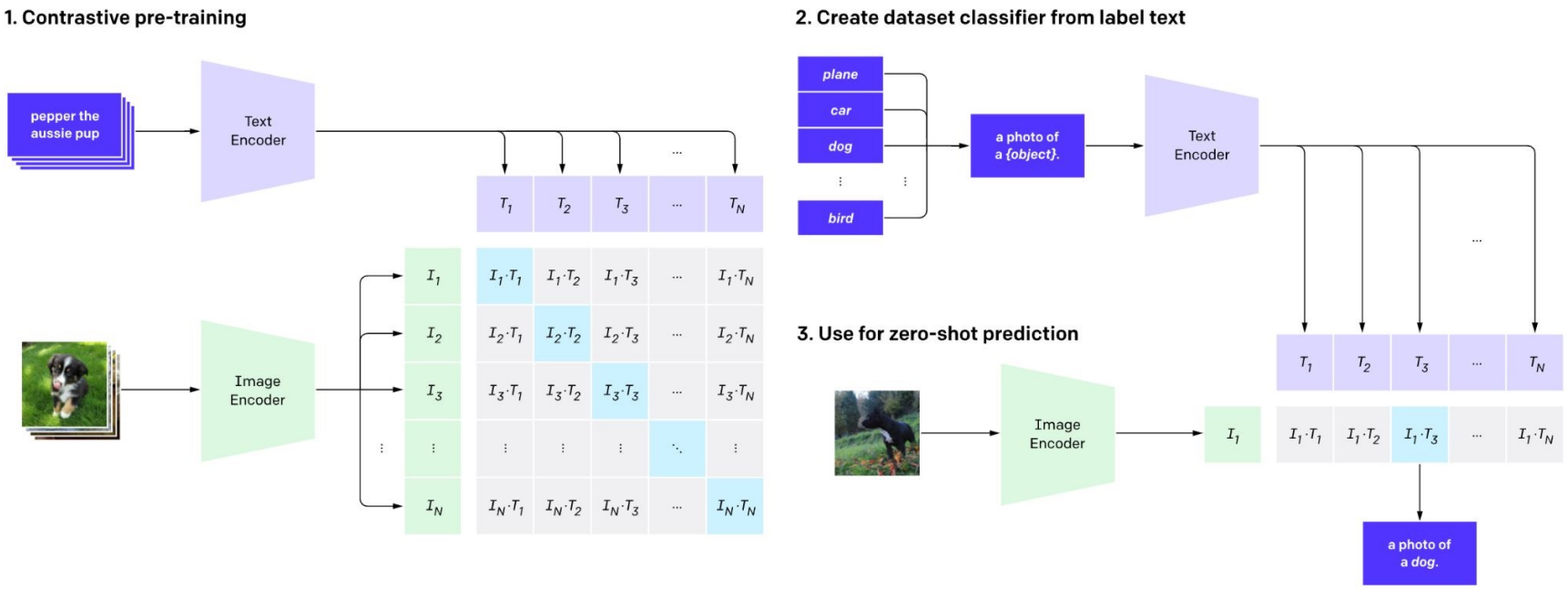
1. What is Natural Language Generation (NLG)?

2. Multimodal Large Language Models (MLLM)

3. Functionality
    - Contrastive Learning Models
    - Generative Models

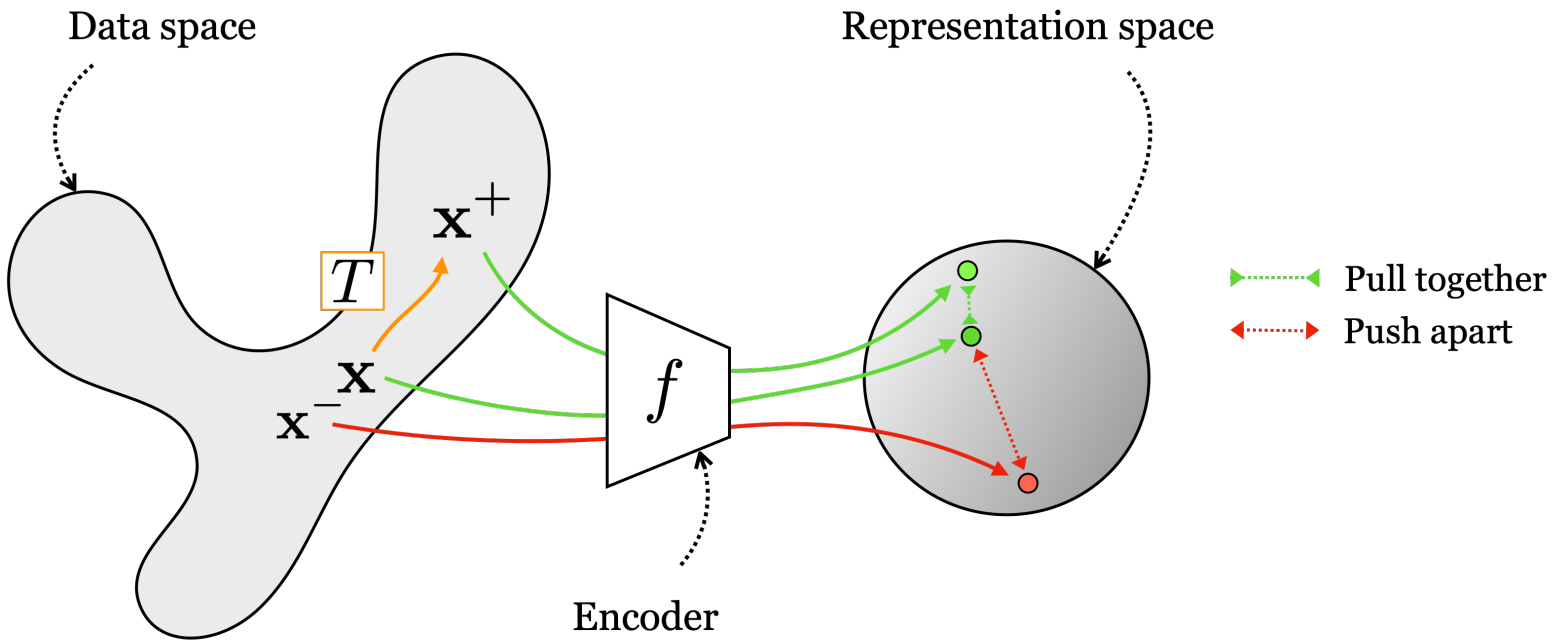
# MLLM Functionality: Image Perception – Contrastive Learning - CLIP (Radford et al. 2021)

Contrastive loss, but.. Transformers and *web data*!




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception – Contrastive Learning

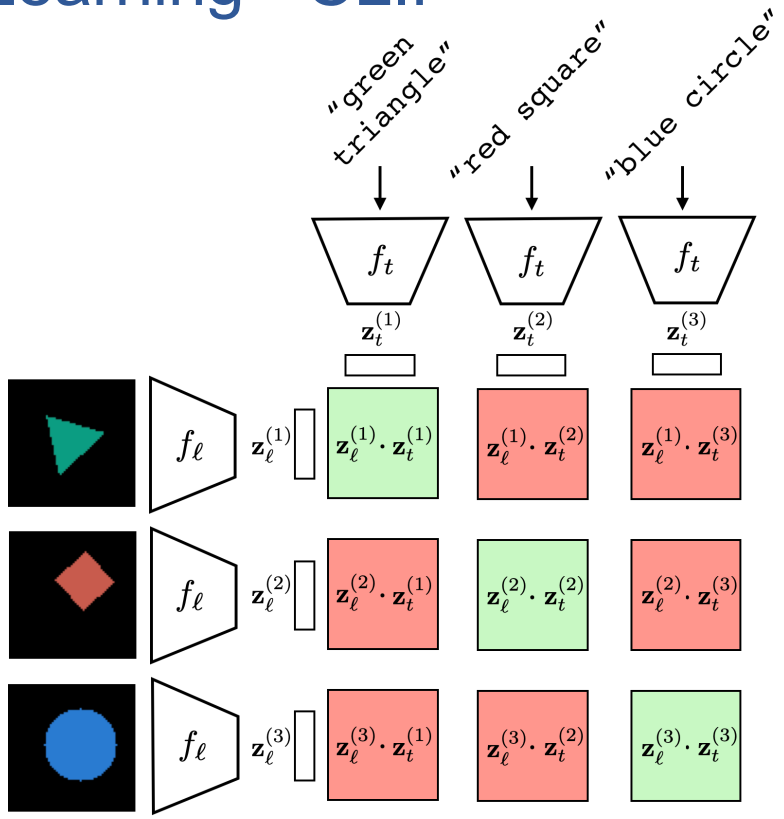


# MLLM Functionality: Image Perception – Contrastive Learning - CLIP

- Image classification: given an image, predict its class name
- Image captioning: given an image, predict its caption
- Contrastive learning: align image and text embeddings that describe the same thing

[Radford*, Kim* et al., ICML 2021]

# MLLM Functionality: Image Perception – Contrastive Learning - CLIP




- Positive pairs: image and its caption.
- Negative pairs: image and a different image's caption.
- Learn a representation in which positives are pulled together, negatives are pushed apart.

[Radford*, Kim* et al., ICML 2021]

# MLLM Functionality: Image Perception – Contrastive Learning - CLIP

**CLIP**

Data

$$\{\mathbf{t}^{(i)}, \boldsymbol{\ell}^{(i)}\}_{i=1}^{N} \rightarrow$$

Objective

$$\left( -\log \frac{e^{f_\ell(\boldsymbol{\ell})^\mathsf{T} f_t(\mathbf{t}^+)/\tau}}{e^{f_\ell(\boldsymbol{\ell})^\mathsf{T} f_t(\mathbf{t}^+)/\tau} + \sum_i e^{f_\ell(\boldsymbol{\ell})^\mathsf{T} f_t(\mathbf{t}_i^-)/\tau}} + \right.$$

$$\left. -\log \frac{e^{f_t(\mathbf{t})^\mathsf{T} f_\ell(\boldsymbol{\ell}^+)/\tau}}{e^{f_t(\mathbf{t})^\mathsf{T} f_\ell(\boldsymbol{\ell}^+)/\tau} + \sum_i e^{f_t(\mathbf{t})^\mathsf{T} f_\ell(\boldsymbol{\ell}_i^-)/\tau}} \right) /2$$

Hypothesis space

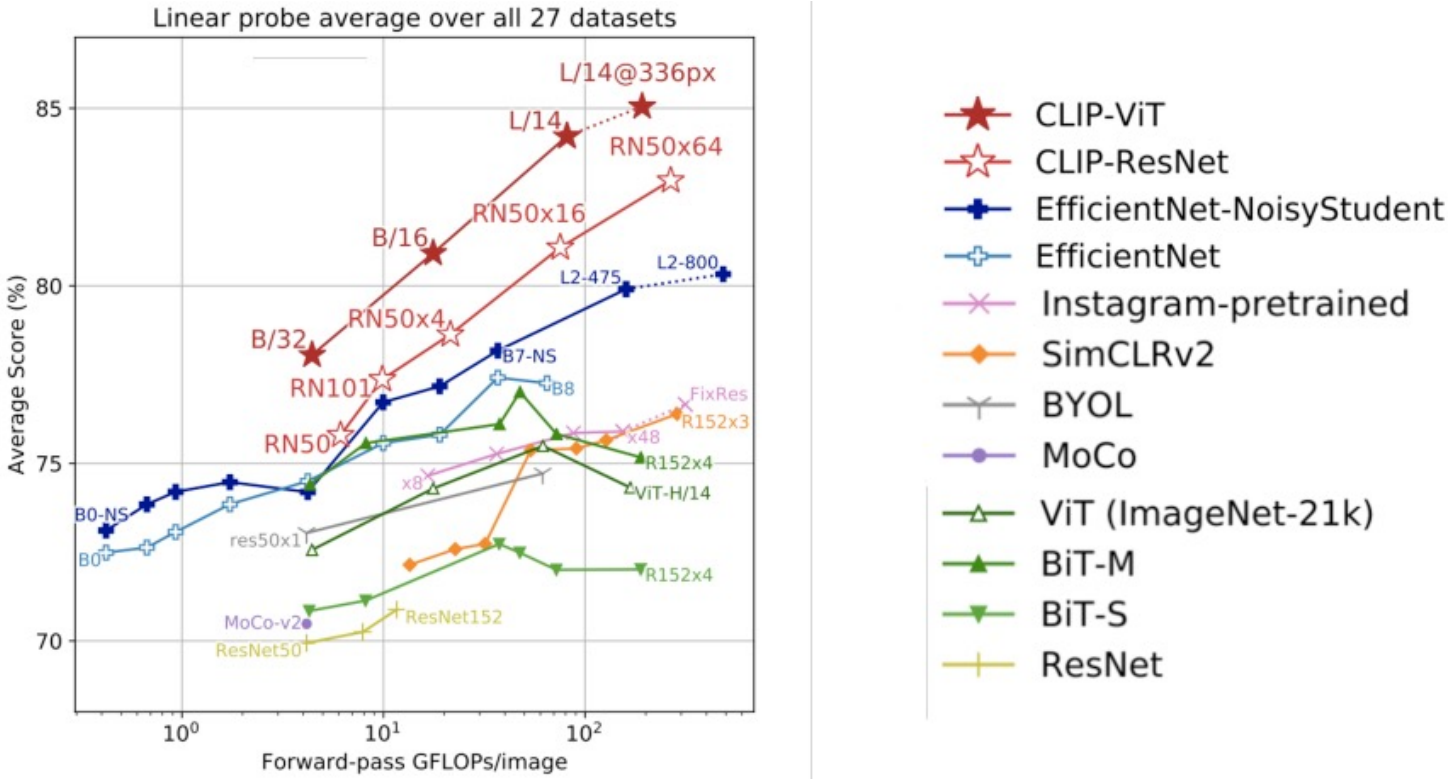
$$f_\ell : \mathbb{R}^{3\times N\times M} \rightarrow \mathbb{S}^{d_z}$$

$$f_t : \mathbb{R}^{T\times d_t} \rightarrow \mathbb{S}^{d_z}$$

$$\rightarrow f_\ell, f_t$$

# MLLM Functionality: Image Perception – Contrastive Learning - CLIP

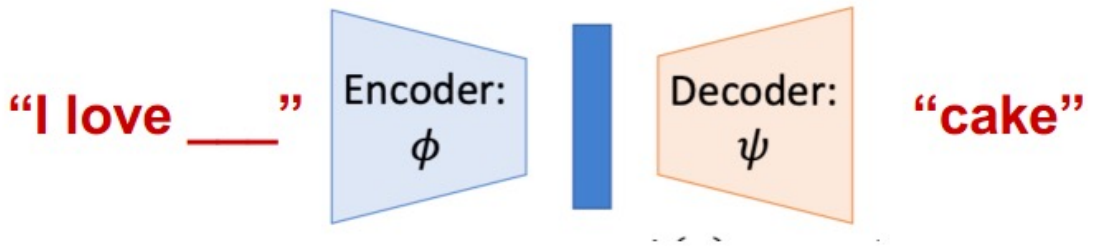
CLIP features w/ **linear probe** across multiple datasets



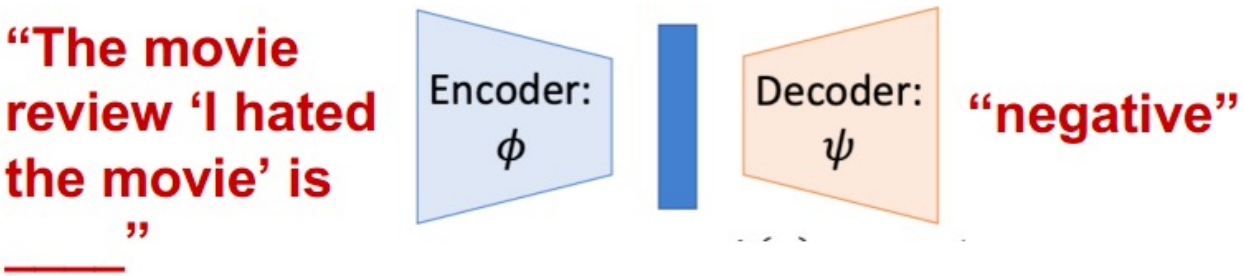

https://cs231n.stanford.edu/

# [CLIP] Big difference with language models: We can use LLMs zero-shot for new downstream tasks

**Step 1:** Pretrain a network on a pretext task that doesn't require supervision



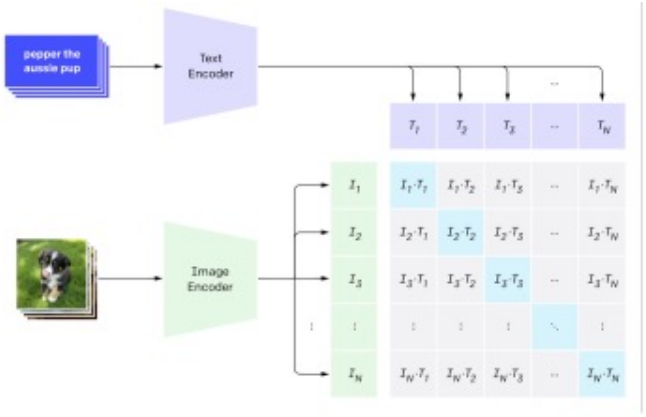

**Step 2:** Use the model out of the box in a creative way!




https://cs231n.stanford.edu/

# [CLIP] But how do we use pre-trained vision-language models in a zero-shot manner?

**Step 1:** Pretrain a network on a pretext task that doesn't require supervision




**Pre-training tasks:**
Contrastive Objective

**Step 2:** Use the model out of the box in a creative way!

**Out of the box classification (No fine-tuning)**

https://cs231n.stanford.edu/

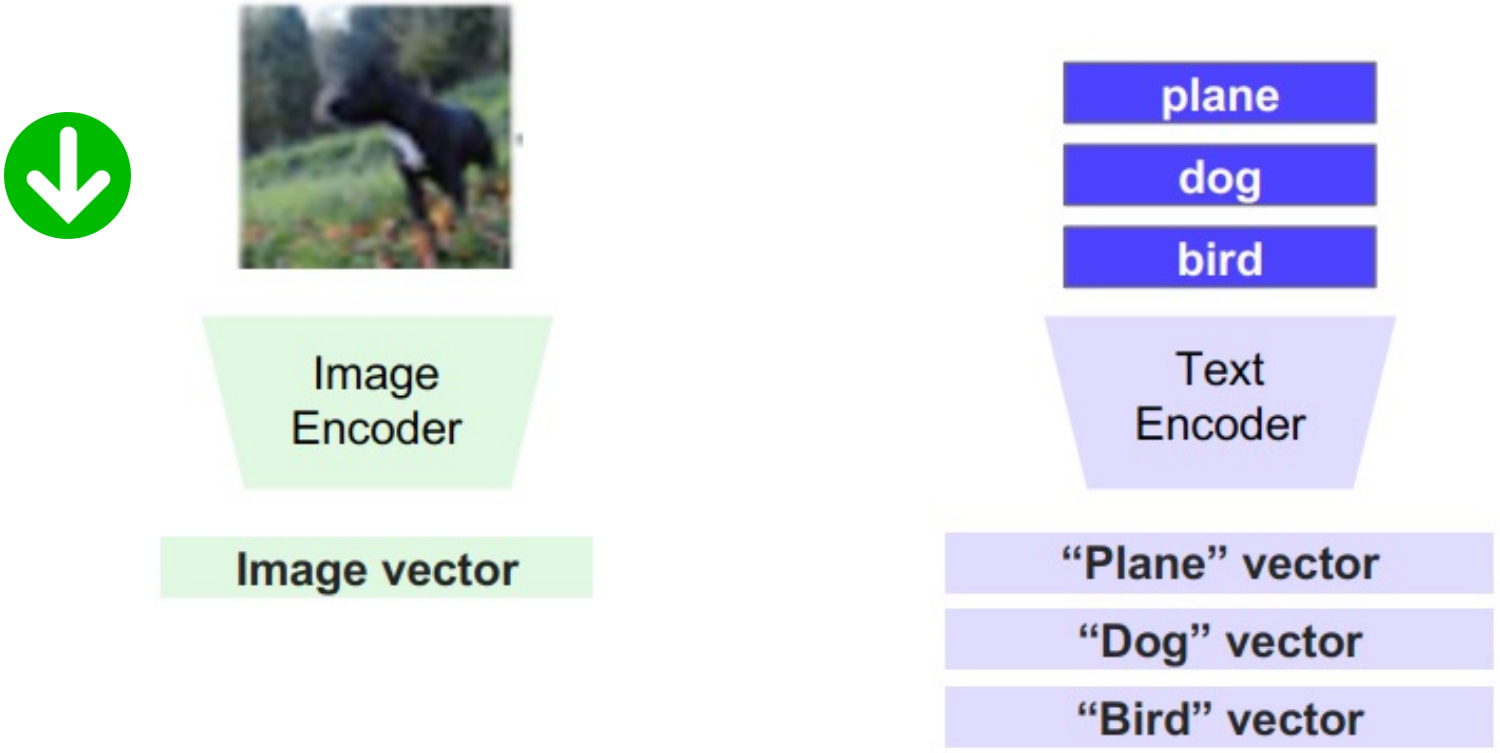
# [CLIP] Clever trick: we can create a classifier using the text encoder!




https://cs231n.stanford.edu/

Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

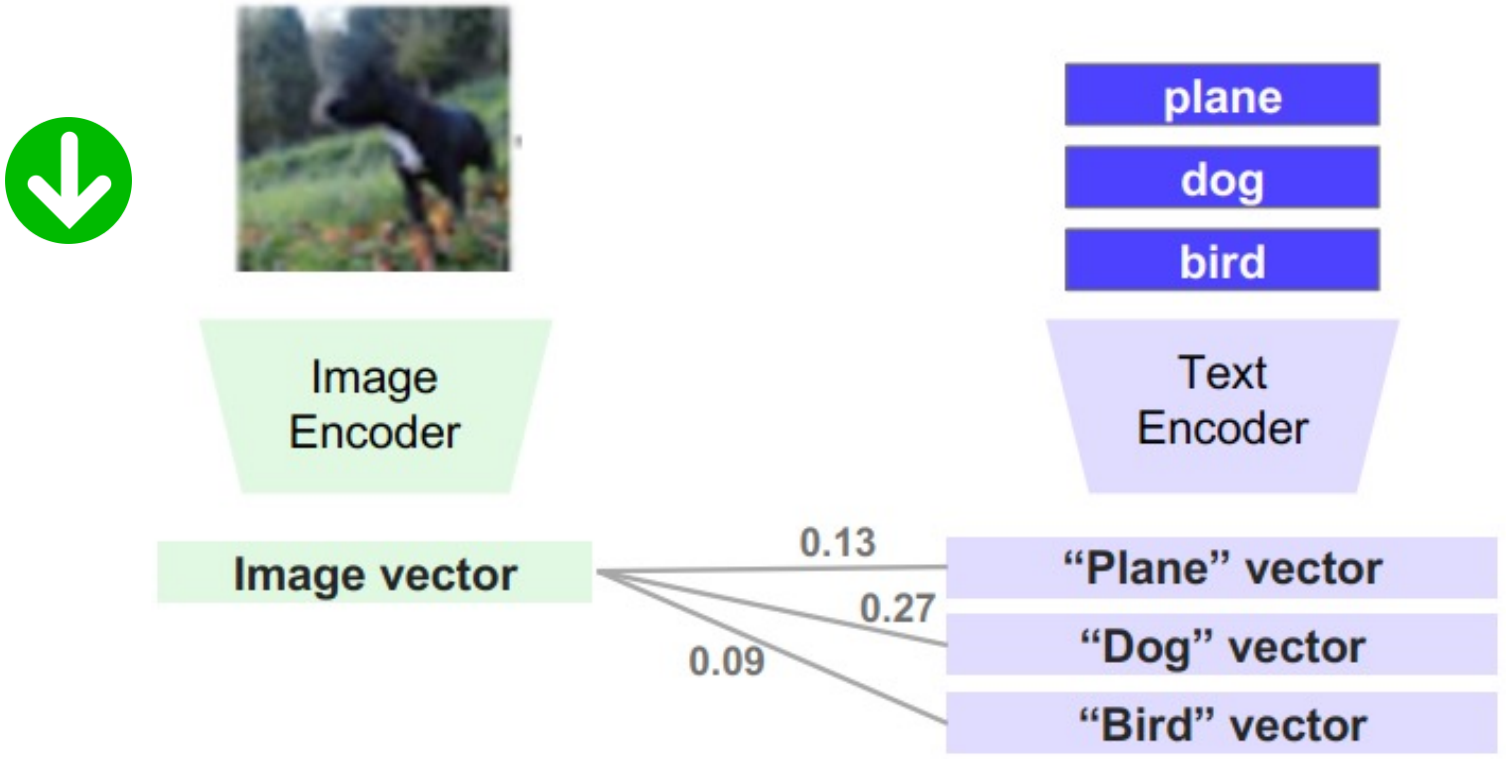
# [CLIP] Create a vector representation for each category!




https://cs231n.stanford.edu/

Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

# [CLIP] Match a new image to the most similar vector




You can think of this as a **1-NN algorithm** with the vectors as the training data

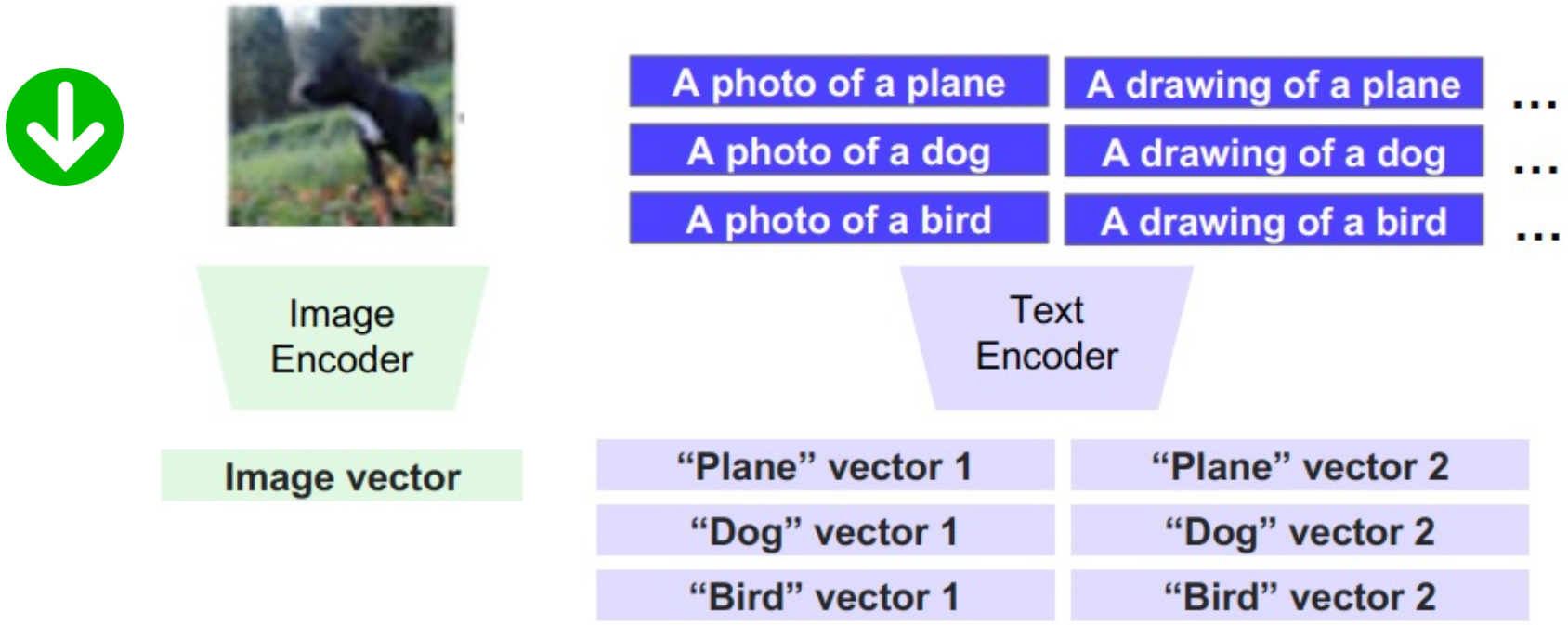
Since CLIP was trained with phrases, you can improve performance by using a phrase “A photo of a [category]”

+1.3% on ImageNet



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

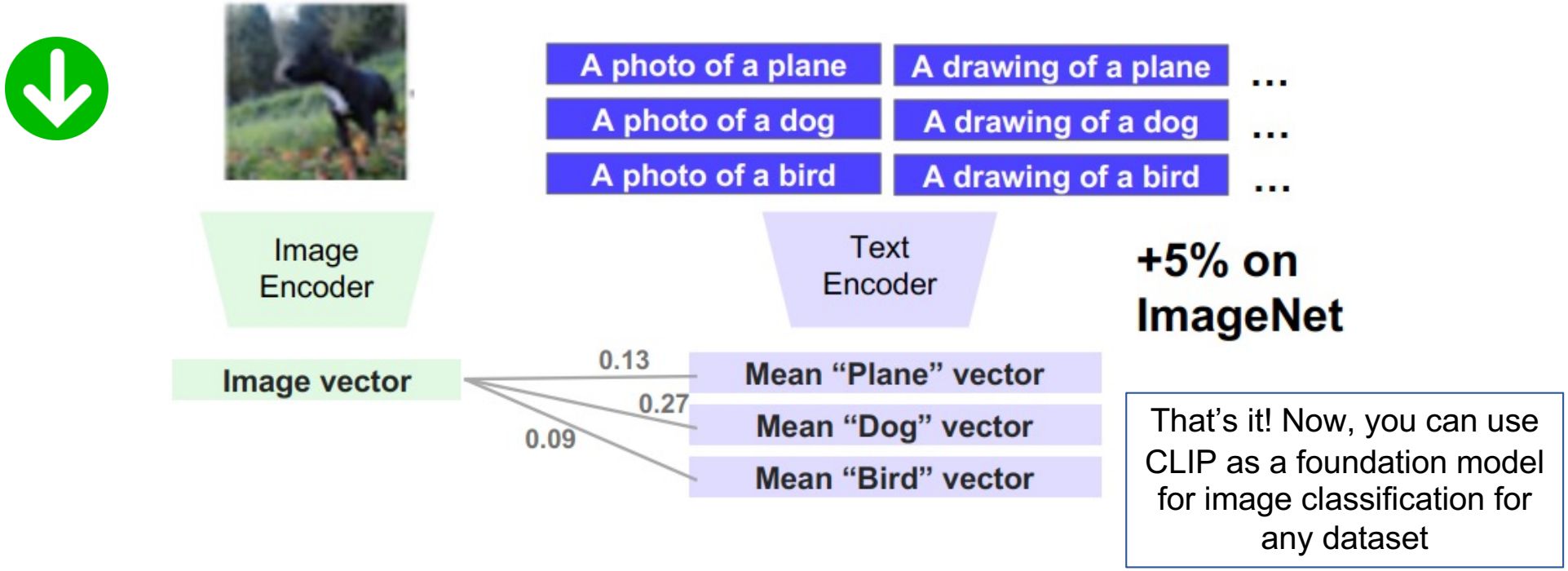
# [CLIP] A single phrase might be too biased. Solution: Use multiple phrases




https://cs231n.stanford.edu/

Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

# [CLIP] Use the average vector across phrases as the representation for each category




That's it! Now, you can use CLIP as a foundation model for image classification for any dataset



Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

# [CLIP] Exciting result after training on 400M image-text pairs

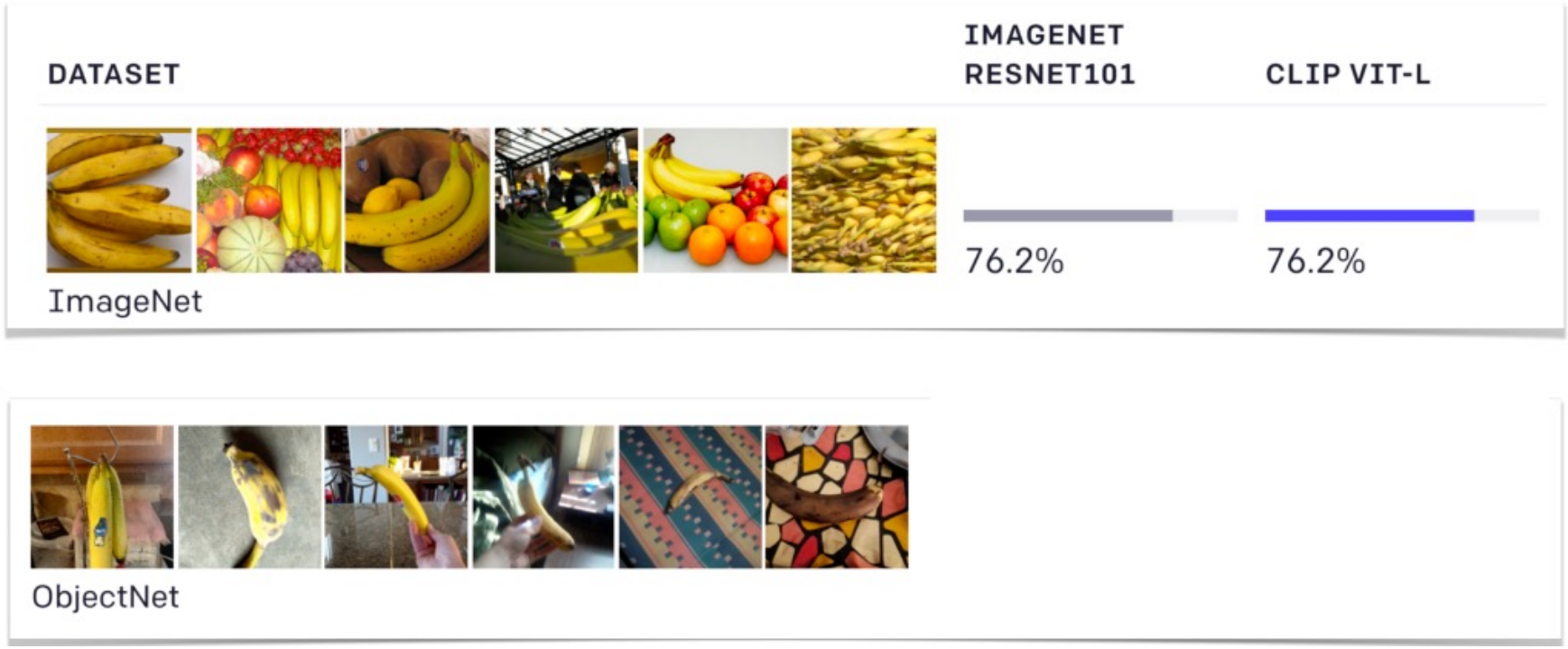



Matches the accuracy of of ResNet 101 that has been trained on ImageNet, except CLIP was trained with no human labels at all!



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

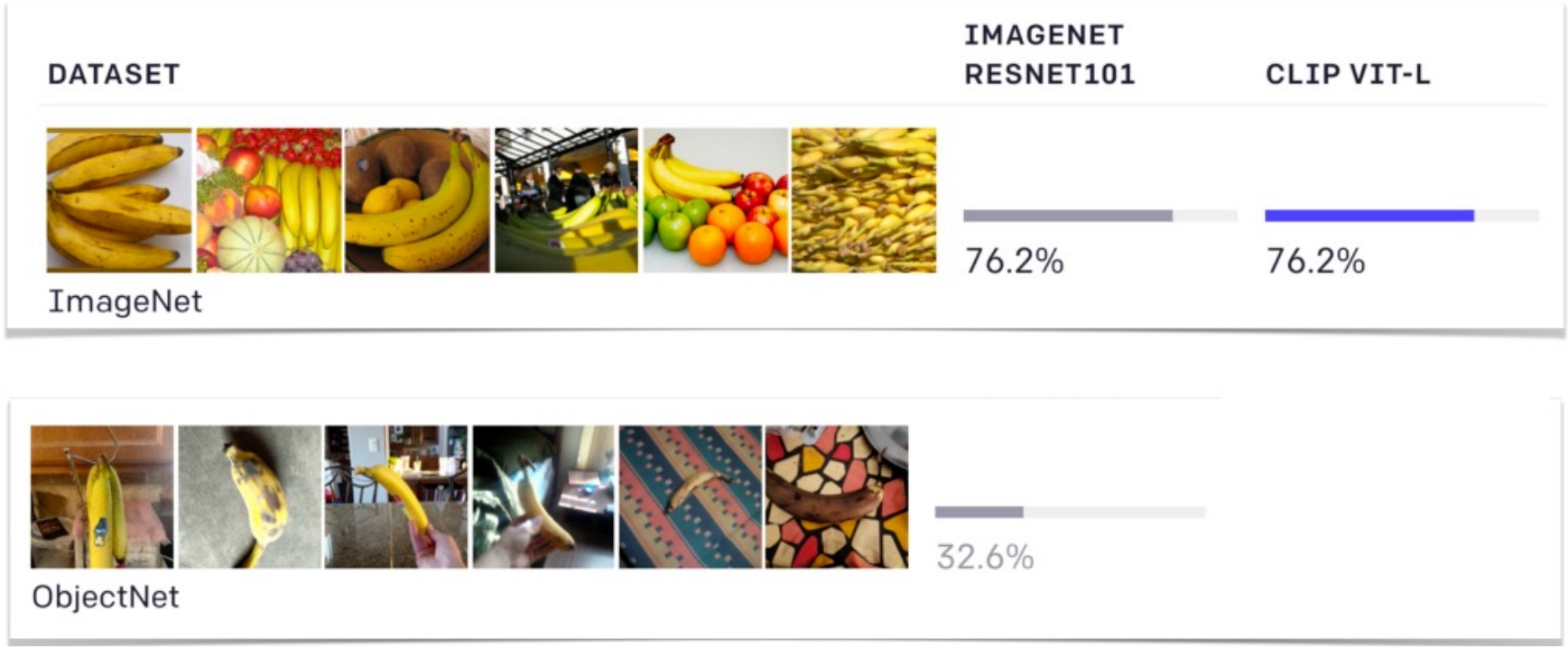
# [CLIP] Here's where things get even more exciting




https://cs231n.stanford.edu/

Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

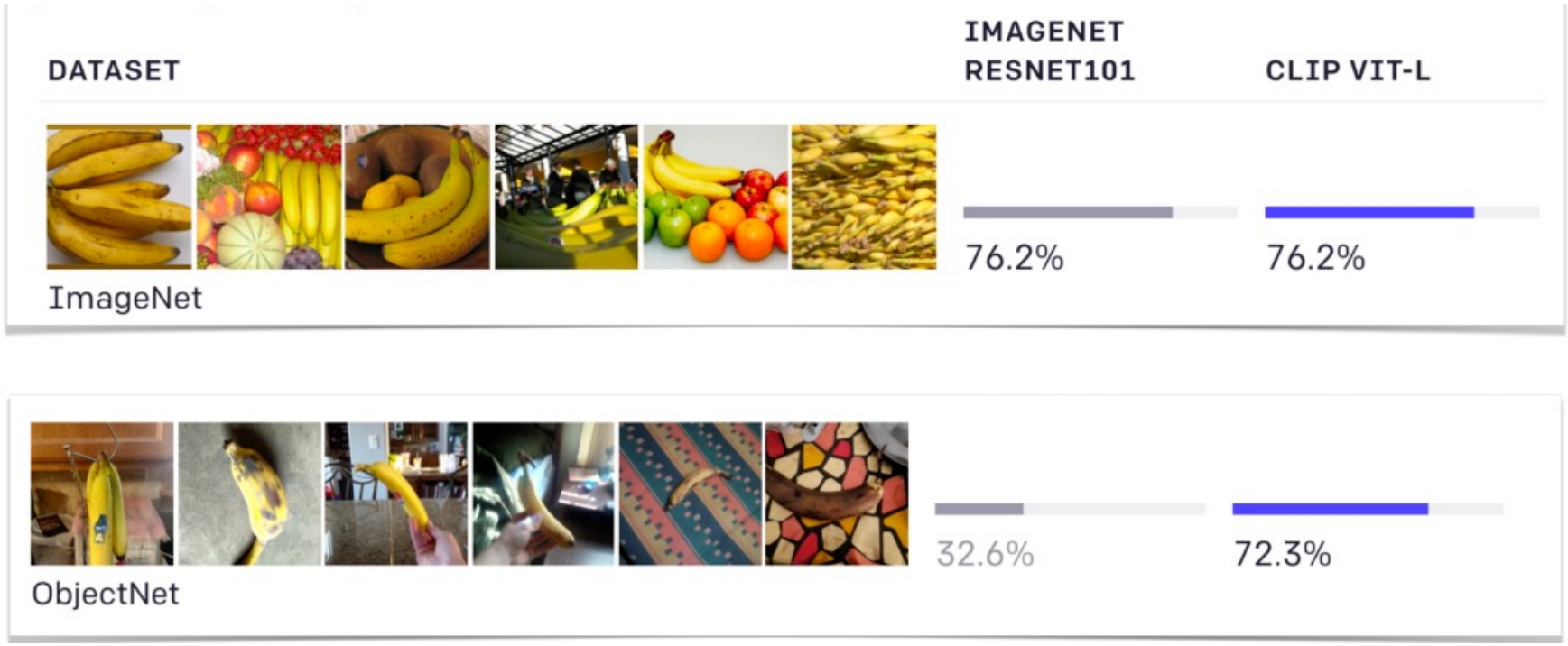
# [CLIP] Training on ImageNet doesn't generalize to other datasets. ObjectNet contains the same categories but in weird viewpoints






Radford et al "Learning Transferable Visual Models From Natural Language Supervision"

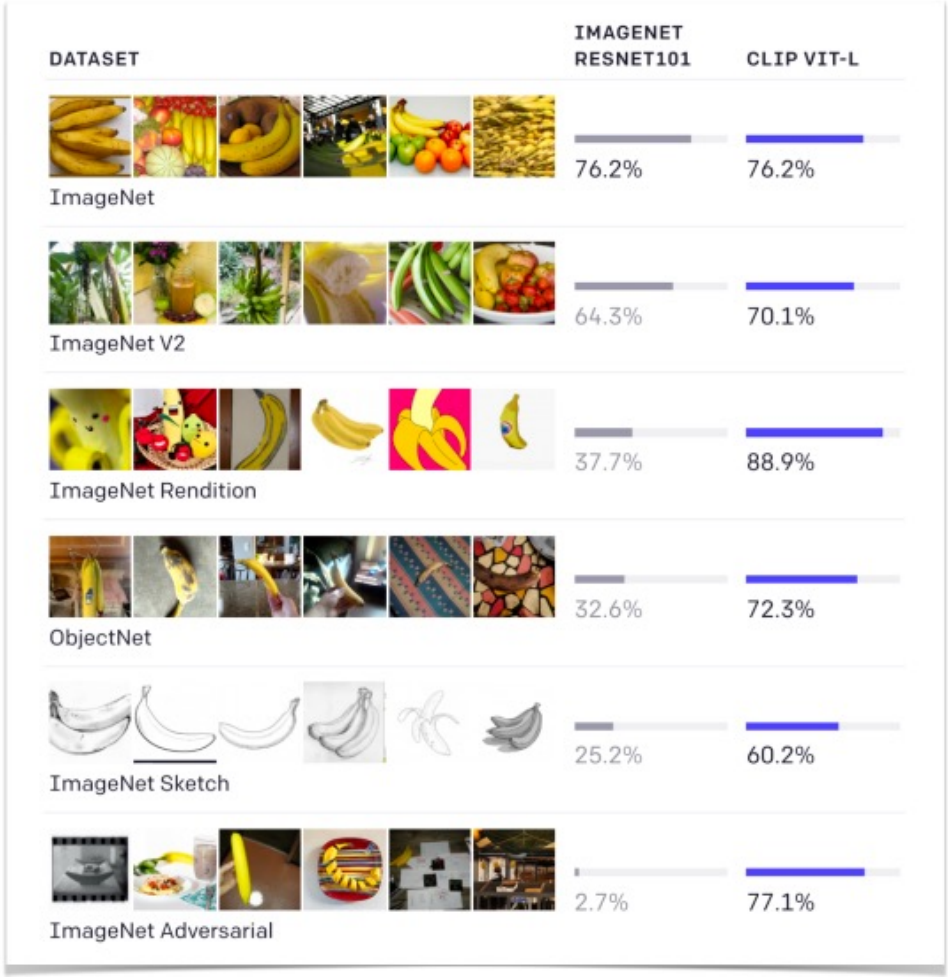
# [CLIP] But CLIP zero-shot does so well! Q. Why do you think that is?






Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

[CLIP] CLIP performance is great also on graphic images, sketches, adversarial datasets, …

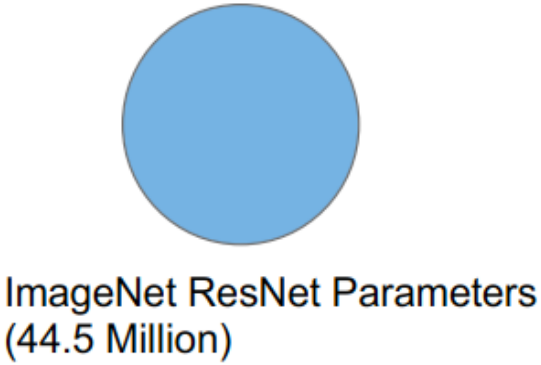


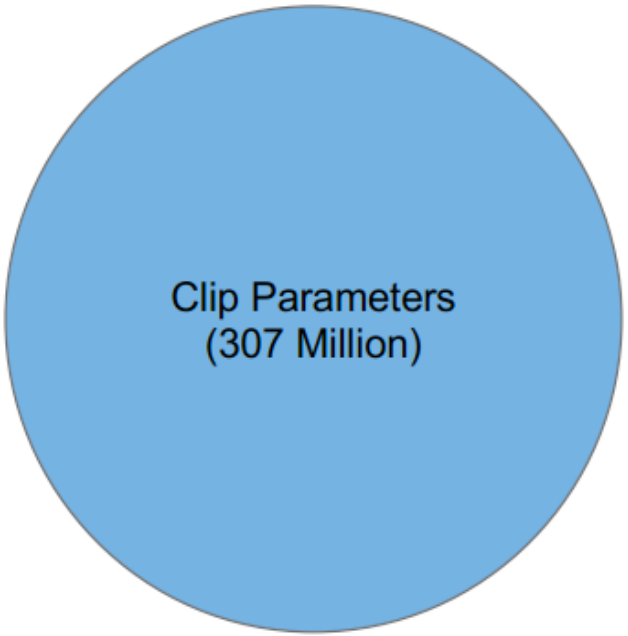



Radford et al “Learning Transferable Visual Models From Natural Language Supervision”

# [CLIP] Why does CLIP perform so well?

## How can no labels beat labels?? Scale!

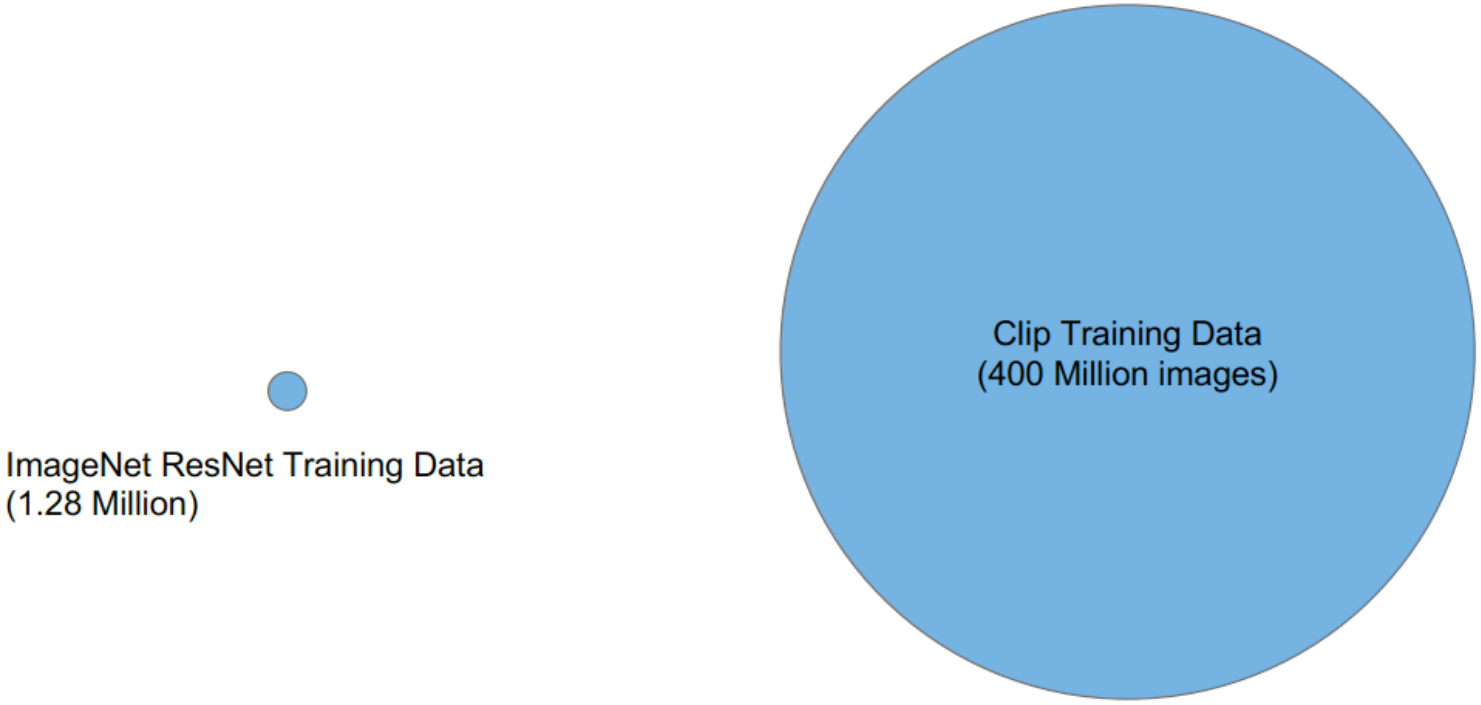



https://cs231n.stanford.edu/

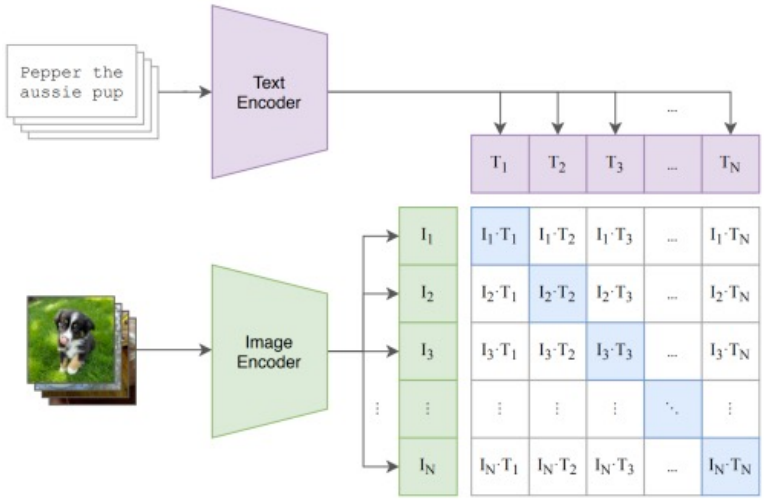
# [CLIP] Why does CLIP perform so well?

## How can no labels beat labels?? Scale!




https://cs231n.stanford.edu/

# SigLIP




**CLIP objective function:**

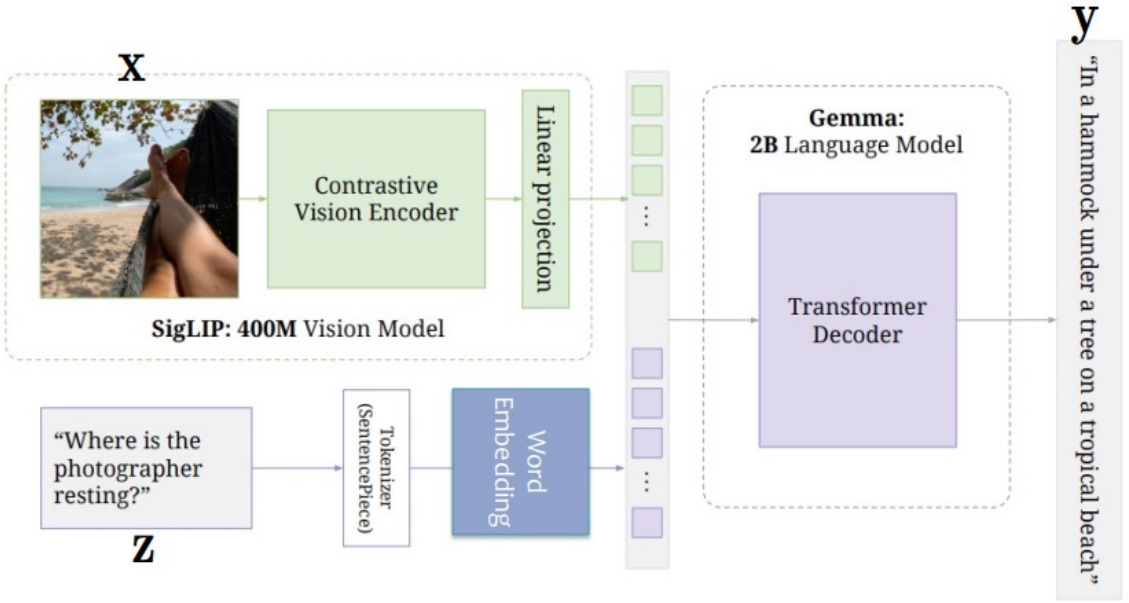
$$\sum_{i=1}^{N}\left[\log\frac{\exp\left(\frac{I_i^\top T_i}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{I_i^\top T_j}{\tau}\right)}+\log\frac{\exp\left(\frac{I_i^\top T_i}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{I_j^\top T_i}{\tau}\right)}\right]$$

https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from Radford et al. (2021) https://arxiv.org/pdf/2103.00020

# PaliGemma

- SigLIP is a variant of CLIP
- Gemma is a 2B LLM (open source counterpart to Gemini)
- Linear projection is creating image embeddings, but **they live in the same high dimensional space as the word embeddings**
- The LLM has to figure out **how to use** the image embeddings in some useful way so as to maximize the likelihood of the correct text response




$$\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{N}$$

https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://arxiv.org/pdf/2407.07726

# MLLM Functionality: Image Perception – Improving CLIP with Descriptions

- Instead of directly querying class category names, can we use descriptions to improve the classification for vision language models?

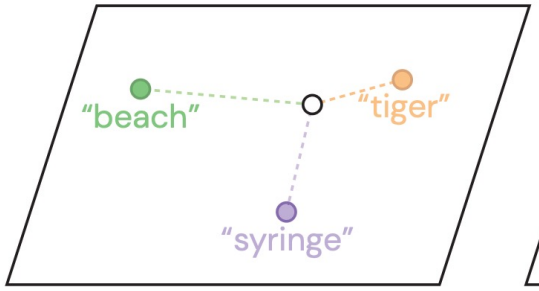
- Can we use language modality as the internal representation for vision task to enhance interpretability?

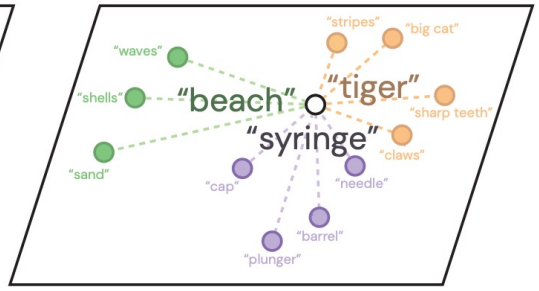


Our top prediction: Hen

and we say that because…

Average
- two legs
- red, brown, or white feathers
- a small body
- a small head
- two wings
- a tail
- a beak
- a chicken



# MLLM Functionality: Image Perception – Visual Classification via Description from Large Language Models (Menon and Vondrick, ICLR 2023)

$$s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x)$$

$s(c, x)$: computed as the addition of all the descriptors pertains to image

- x: image
- d: descriptor
- D(c): descriptors for class c
- phi: dot product using CLIP

```
for i, (k, v) in enumerate(description_encodings.items()): # You can also vectorize this; it wasn't much

    dot_product_matrix = image_encodings @ v.T

    image_description_similarity[i] = dot_product_matrix
    image_description_similarity_cumulative[i] = aggregate_similarity(image_description_similarity[i])
```

# MLLM Functionality: Image Perception – Visual Classification via Description from Large Language Models (Menon and Vondrick, ICLR 2023)

## Generating Descriptors from Large Language Models (LLMs)

```
Q: What are useful features for distinguishing a {category
    name} in a photo?
A: There are several useful visual features to tell there is a
    {category name} in a photo:
-
```

Adding "-" help elicit LLMs to output in a bulleted list

Prompt Structure

```
Q: What are useful visual features for distinguishing a lemur in a photo?
A: There are several useful visual features to tell there is a lemur in a
    photo:
- four-limbed primate
- black, grey, white, brown, or red-brown
- wet and hairless nose with curved nostrils
- long tail
- large eyes
- furry bodies
- clawed hands and feet
```

Haolin Zhang; Menon and Vondrick, ICLR 2023

# MLLM Functionality: Image Perception – Visual Classification via Description from Large Language Models (Menon and Vondrick, ICLR 2023)

## Classification by Description

$$\arg\max_{c \in C} s(c, x)$$

$$s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x)$$

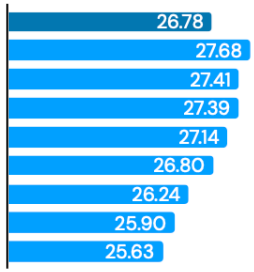
Computed by CLIP similarity



Our top prediction: Hen

and we say that because...

Average
- two legs
- red, brown, or white feathers
- a small body
- a small head
- two wings
- a tail
- a beak
- a chicken



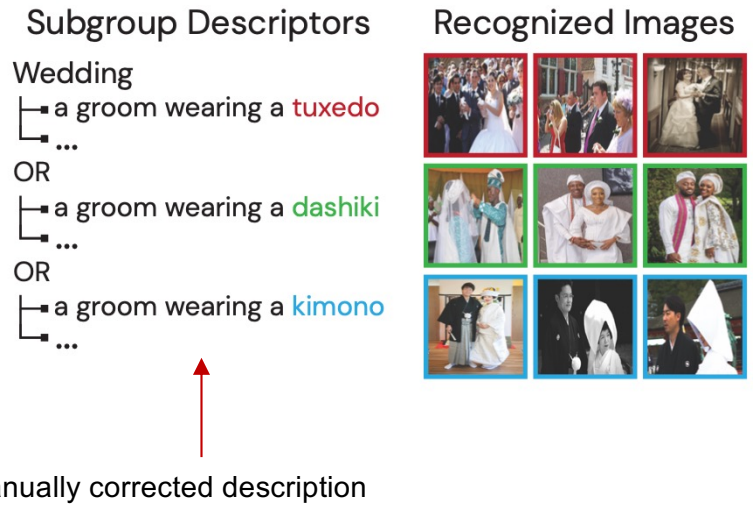

CLIP's top prediction: Dalmatian

Haolin Zhang; Menon and Vondrick, ICLR 2023

# MLLM Functionality: Image Perception – Visual Classification via Description from Large Language Models (Menon and Vondrick, ICLR 2023)

## Correcting failures induced by bias

- Both foundational models (CLIP and GPT 3) have bias for certain categories ---- e.g. "Wedding"




| Sub-group | Ours | CLIP |
|---|---|---|
| Western African | 100% | 40% |
| Chinese | 100% | 20% |
| Japanese | 100% | 0% |
| North Indian | 100% | 60% |

Haolin Zhang; Menon and Vondrick, ICLR 2023

# MLLM Functionality: Image Perception – Visual Classification via Description from Large Language Models (Menon and Vondrick, ICLR 2023)

## Classification by Description (Results)

### Analyzing the failure modes

- Failure in descriptor creation



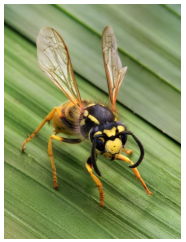
Vespa from ImageNet

Jackfruit, which (has/is/etc)
- large, round fruit
- green or yellow skin
- white flesh with black seeds
- sweet and sticky taste
- strong smell

Not visual descriptors

Vespa, which (has/is/etc)
- a type of wasp
- black and yellow stripes
- a stinger
- two pairs of wings
- six legs
- a narrow waist

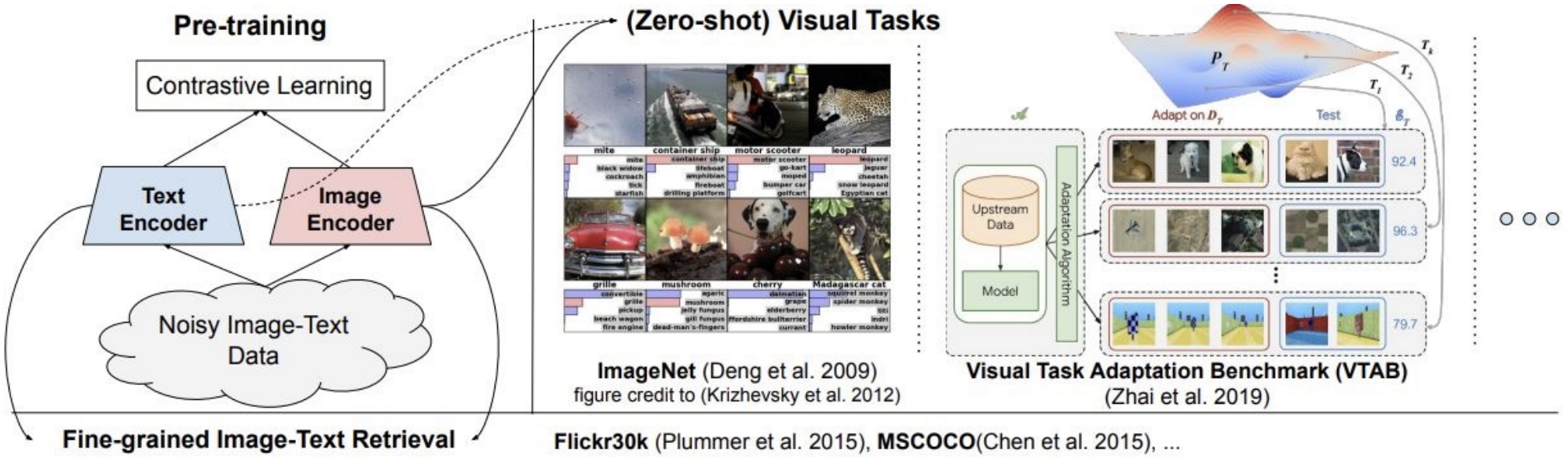
From GPT 3



Hair spray, which (has/is/etc)
- aerosable product
- aerosable product
- aerosable product

# MLLM Functionality: Image Perception – Contrastive Learning - ALIGN (Jia et al., 2021)

Same idea of CLIP, but EVEN MORE data (JFT at 1.8B image-text pairs vs CLIP's 300m).



# MLLM Functionality: Image Perception

FLAVA (Singh et al., 2021)



Holistic approach to multimodality.
One foundation model spanning V&L, CV and NLP.

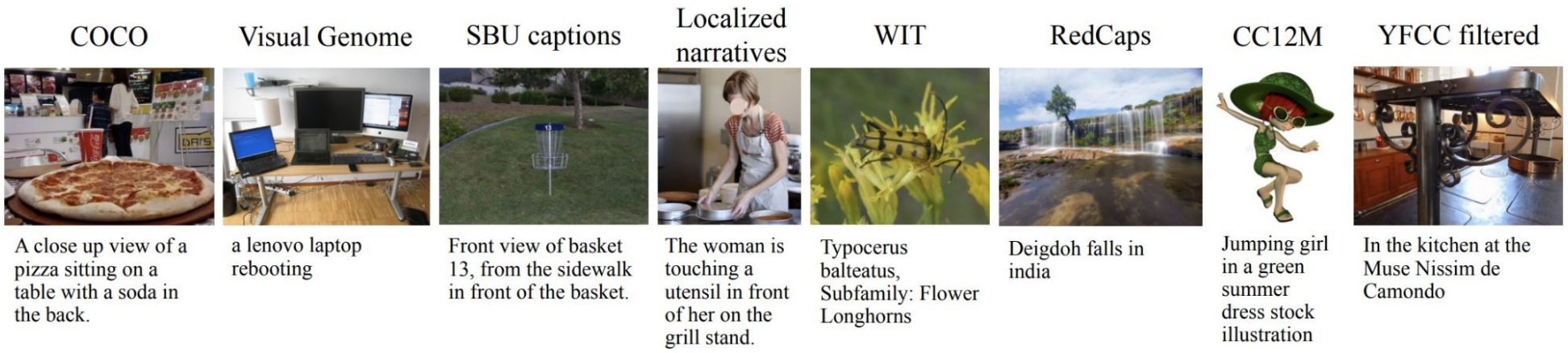
Jointly pretrained on:
- unimodal text data (CCNews + BookCorpus)
- unimodal image data (ImageNet)
- public paired image-text data (70M)

All data/models are publicly released.

# MLLM Functionality: Image Perception - FLAVA

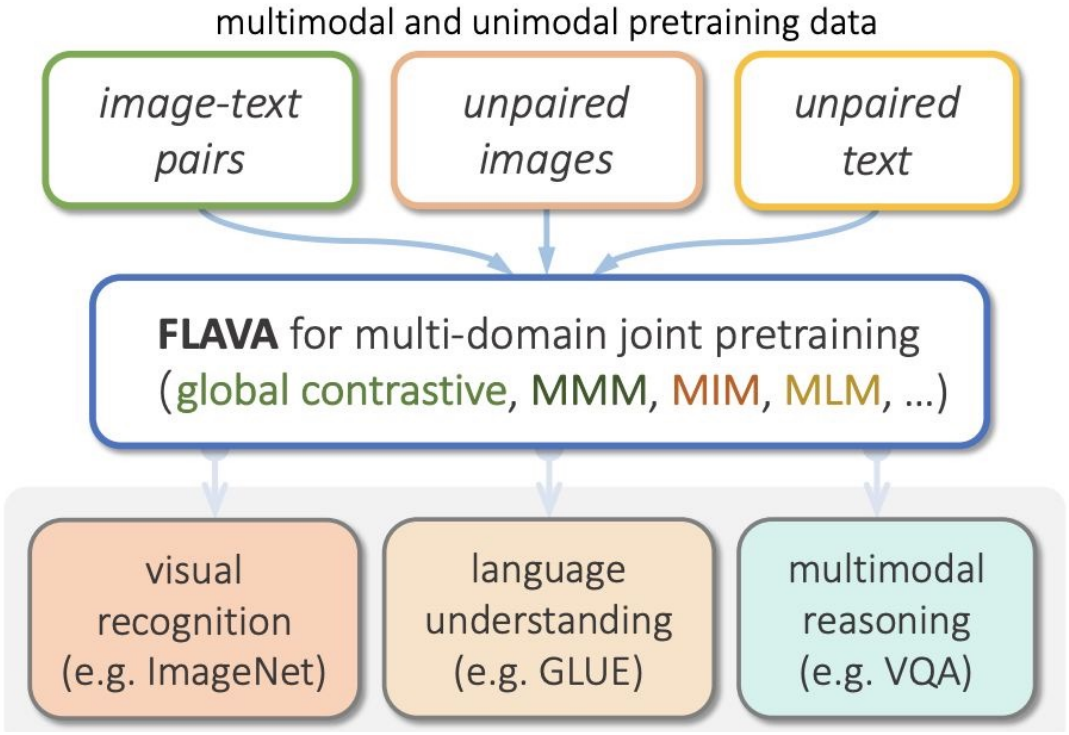
## The PMD dataset

- 70M image-text pairs from public sources



# MLLM Functionality: Image Perception - FLAVA

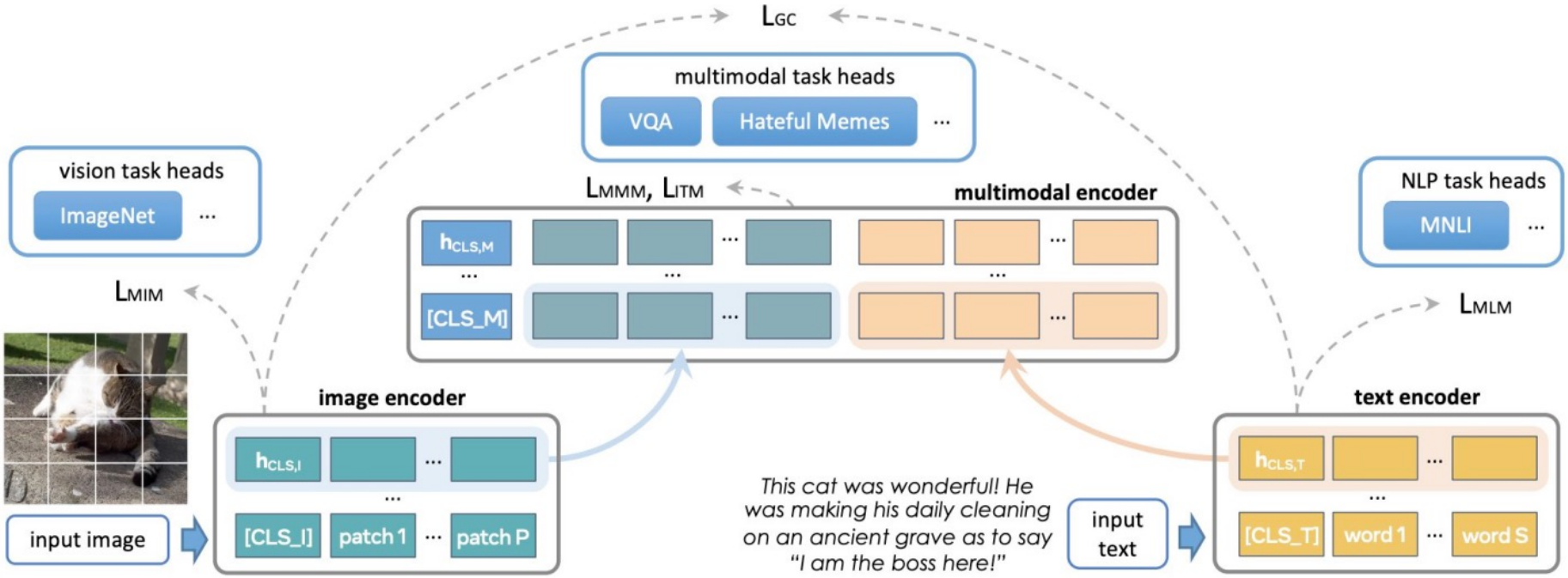
## Problems to Solve




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception - FLAVA

How does FLAVA work?




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# Plan for this lecture

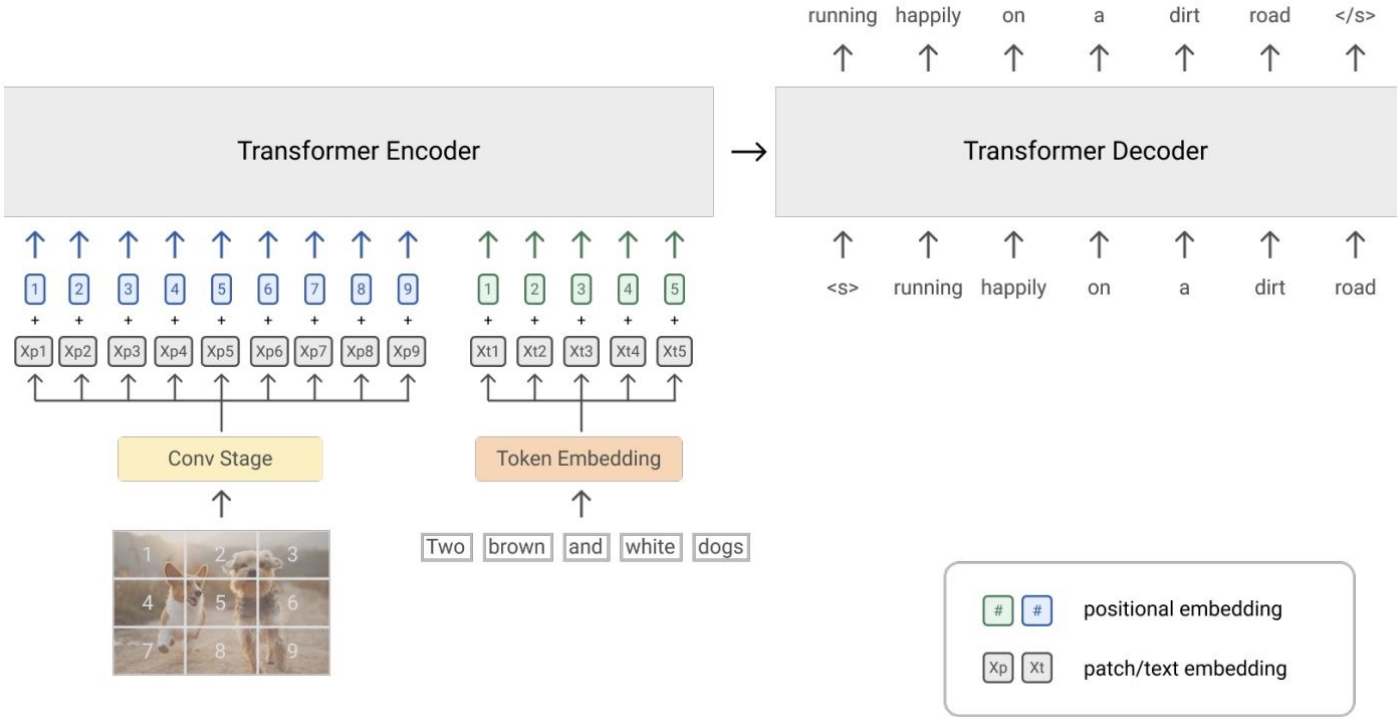
1. What is Natural Language Generation (NLG)?

2. Multimodal Large Language Models (MLLM)

3. Functionality
   - Contrastive Learning Models
   - Generative Models

# MLLM Functionality: Image Perception - SimVLM (Wang et al., 2022)

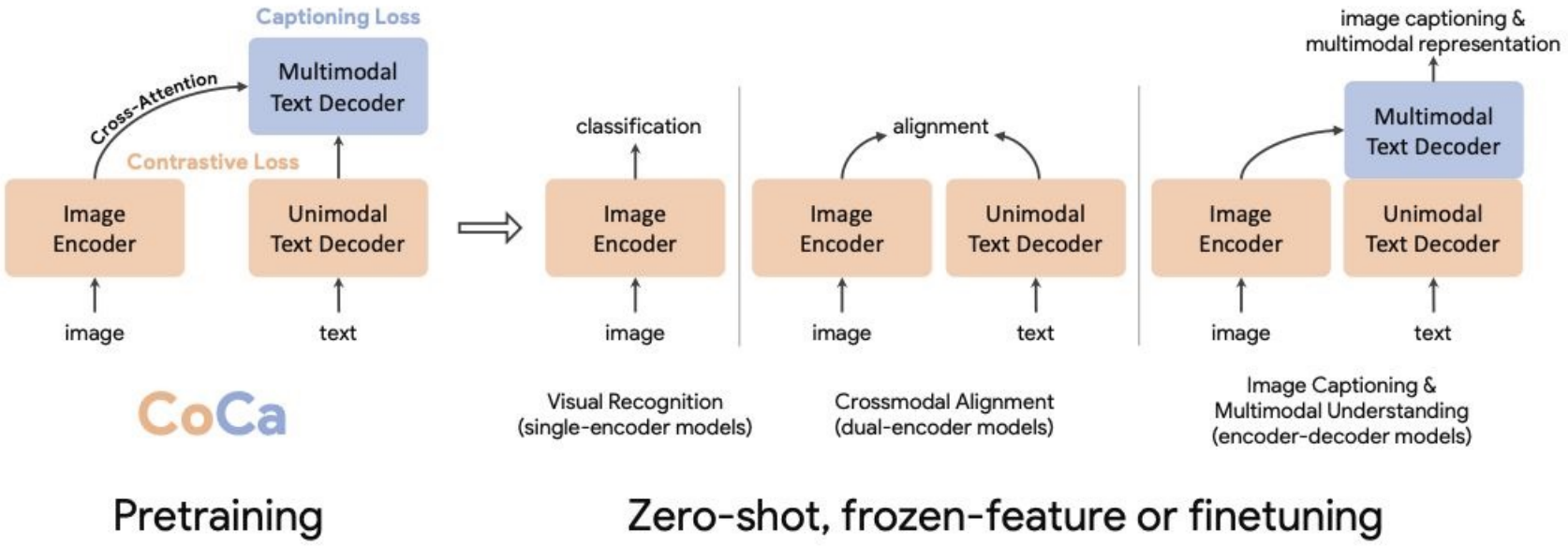
Slowly moving from
contrastive/discriminative to generative.




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception - CoCa Contrastive Captioner (Yu et al., 2022)

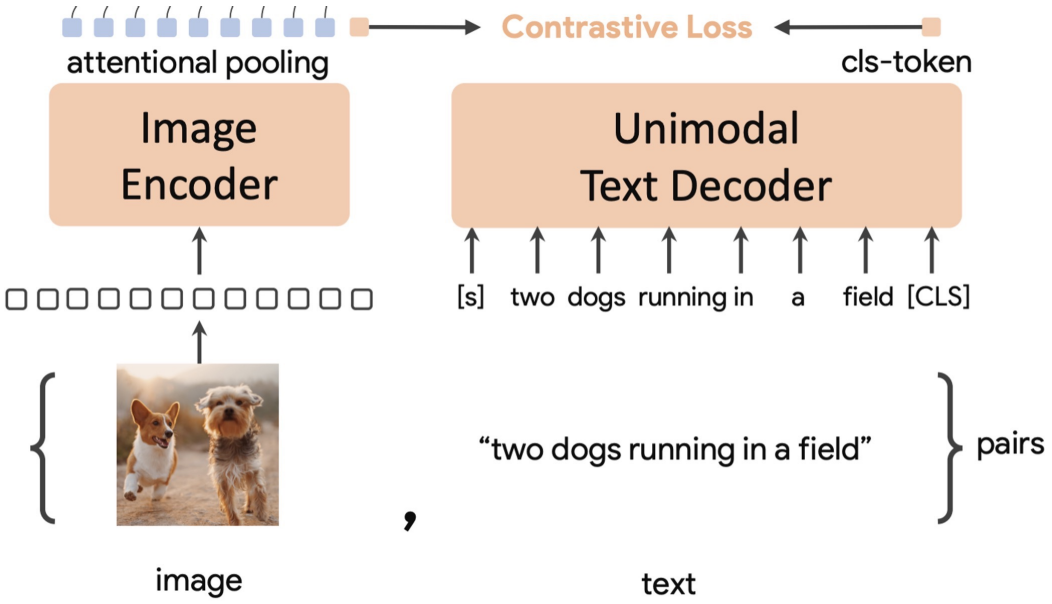
Best of both (contrastive and generative) worlds.




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception - CoCa Contrastive Captioner (Yu et al., 2022)

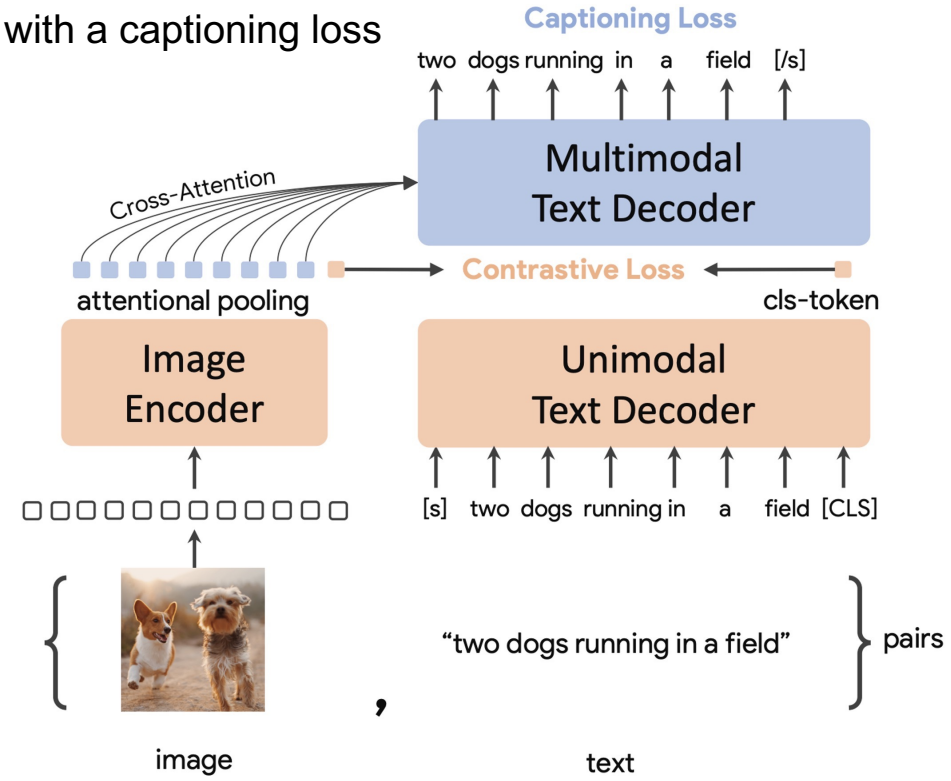
CoCa improved upon CLIP by adding a generation objective




"Contrastive Captioners are Image-Text Foundation Models", 2022

# MLLM Functionality: Image Perception - CoCa Contrastive Captioner (Yu et al., 2022)

CoCa added a decoder with a captioning loss




"Contrastive Captioners are Image-Text Foundation Models", 2022

# MLLM Functionality: Image Perception - CoCa Contrastive Captioner (Yu et al., 2022)

CoCa: Contrastive Captioners are Image-Text Foundation Models

| Model | ImageNet | ImageNet-A | ImageNet-R | ImageNet-V2 | ImageNet-Sketch | ObjectNet | Average |
|---|---|---|---|---|---|---|---|
| CLIP [12] | 76.2 | 77.2 | 88.9 | 70.1 | 60.2 | 72.3 | 74.3 |
| ALIGN [13] | 76.4 | 75.8 | 92.2 | 70.1 | 64.8 | 72.2 | 74.5 |
| FILIP [61] | 78.3 | - | - | - | - | - | - |
| Florence [14] | 83.7 | - | - | - | - | - | - |
| LiT [32] | 84.5 | 79.4 | 93.9 | 78.7 | - | 81.1 | - |
| BASIC [33] | 85.7 | 85.6 | 95.7 | 80.6 | 76.1 | 78.9 | 83.7 |
| CoCa-Base | 82.6 | 76.4 | 93.2 | 76.5 | 71.7 | 71.6 | 78.7 |
| CoCa-Large | 84.8 | 85.7 | 95.6 | 79.6 | 75.7 | 78.6 | 83.3 |
| CoCa | **86.3** | **90.2** | **96.5** | **80.7** | **77.6** | **82.7** | **85.7** |

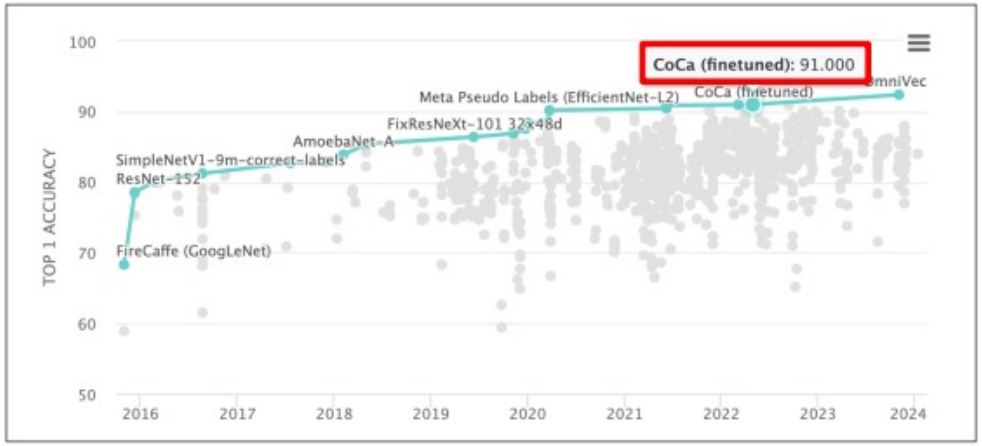
Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

"Contrastive Captioners are Image-Text Foundation Models", 2022

# MLLM Functionality: Image Perception - CoCa Contrastive Captioner (Yu et al., 2022)

Classifier foundation models now beat all other models on ImageNet

| Model | ImageNet |
|---|---|
| ALIGN [13] | 88.6 |
| Florence [14] | 90.1 |
| MetaPseudoLabels [51] | 90.2 |
| CoAtNet [10] | 90.9 |
| ViT-G [21] | 90.5 |
| + Model Soups [52] | 90.9 |
| CoCa (frozen) | 90.6 |
| CoCa (finetuned) | **91.0** |

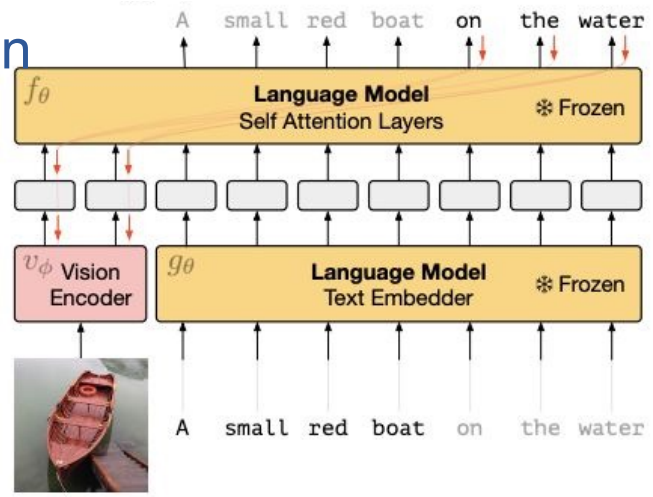


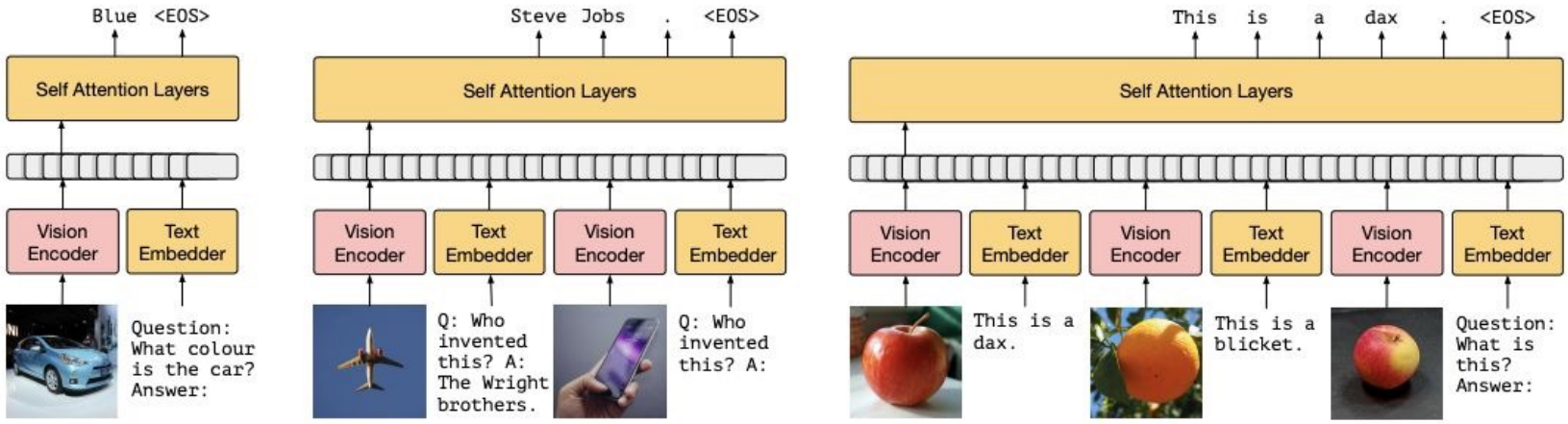

"Contrastive Captioners are Image-Text Foundation Models", 2022

# MLLM Functionality: Image Perception - Frozen (Tsimpoukelli, Menick, Cabi, et al., 2021)

Kind of like Multimodal BiTransformers (MMBT) but with a better LLM (T5) and a better vision encoder (NF-ResNet).

Multi-Modal Few-Shot Learners!




(a) **0-shot VQA** (b) **1-shot outside-knowledge VQA** (c) **Few-shot image classification**

Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

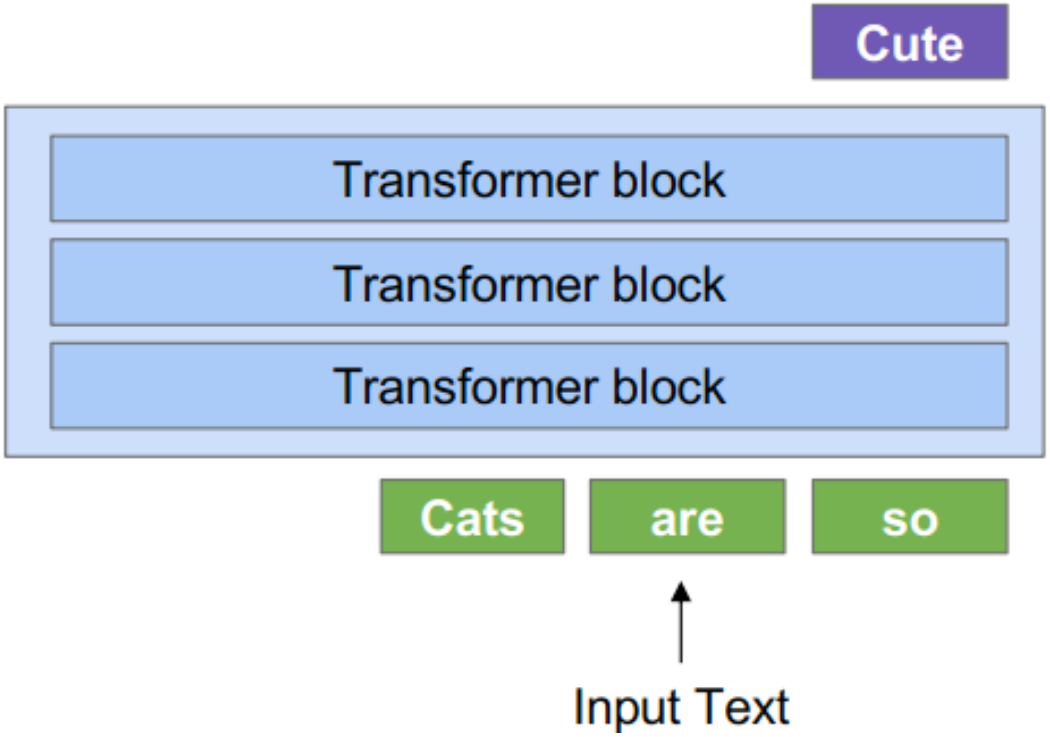
# MLLM Functionality: Image Perception - LLaVA: Large Language and Vision Assistant (Liu NeurIPS'23)

- Motivation: Language models which do next token prediction can be applied to a wide variety of tasks at inference (Math, sentiment analysis, symbolic reasoning)

- Can we build a model that can accept images and text as input, and then output text?

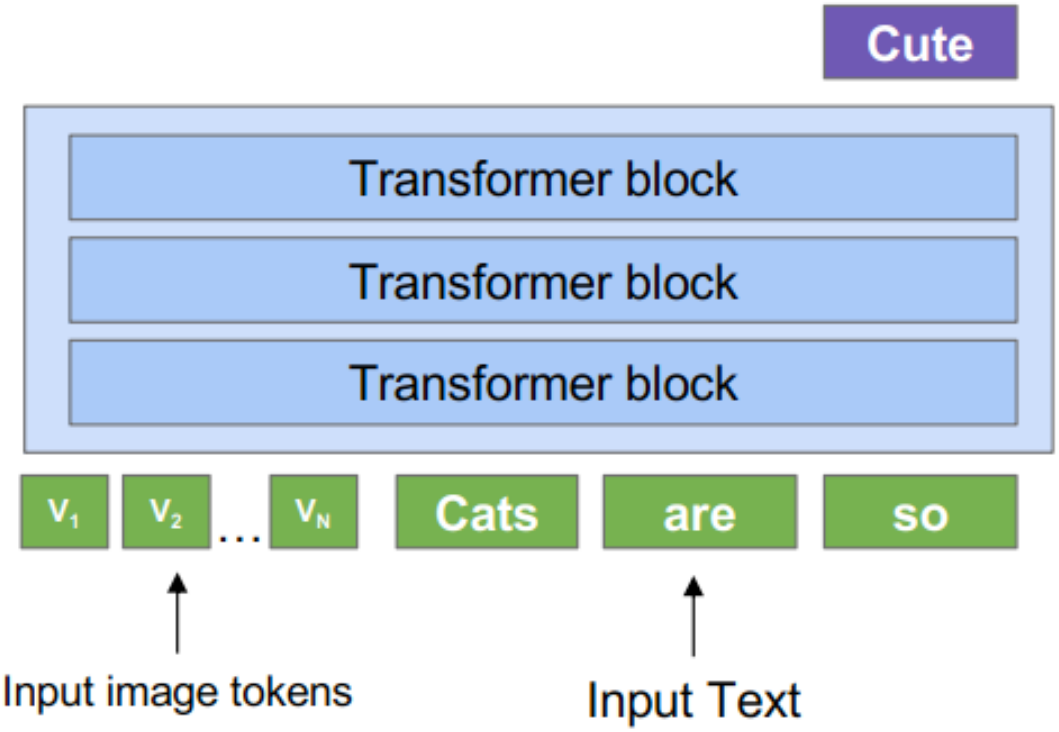
- → Vision-Language Models

# LLaVA: Recall how transformers decode language




https://cs231n.stanford.edu/

# Key idea behind LLaVA – add visual information to the LLM

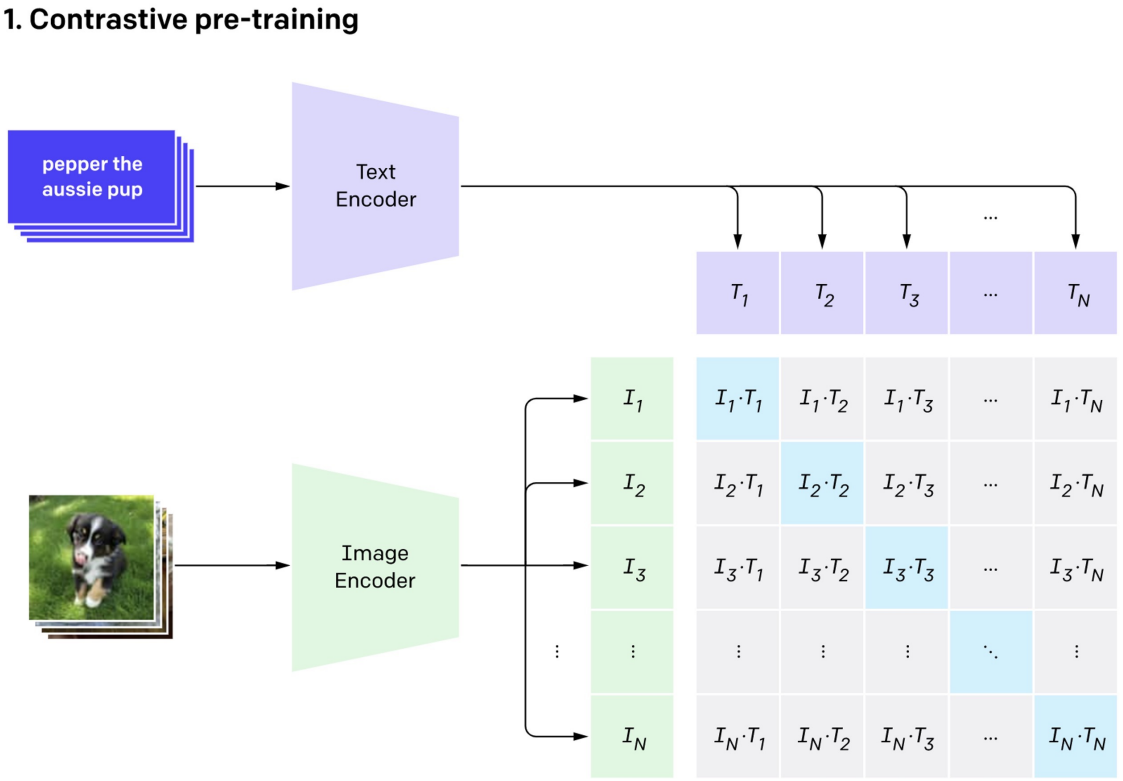



Which image tokens work best here?

https://cs231n.stanford.edu/

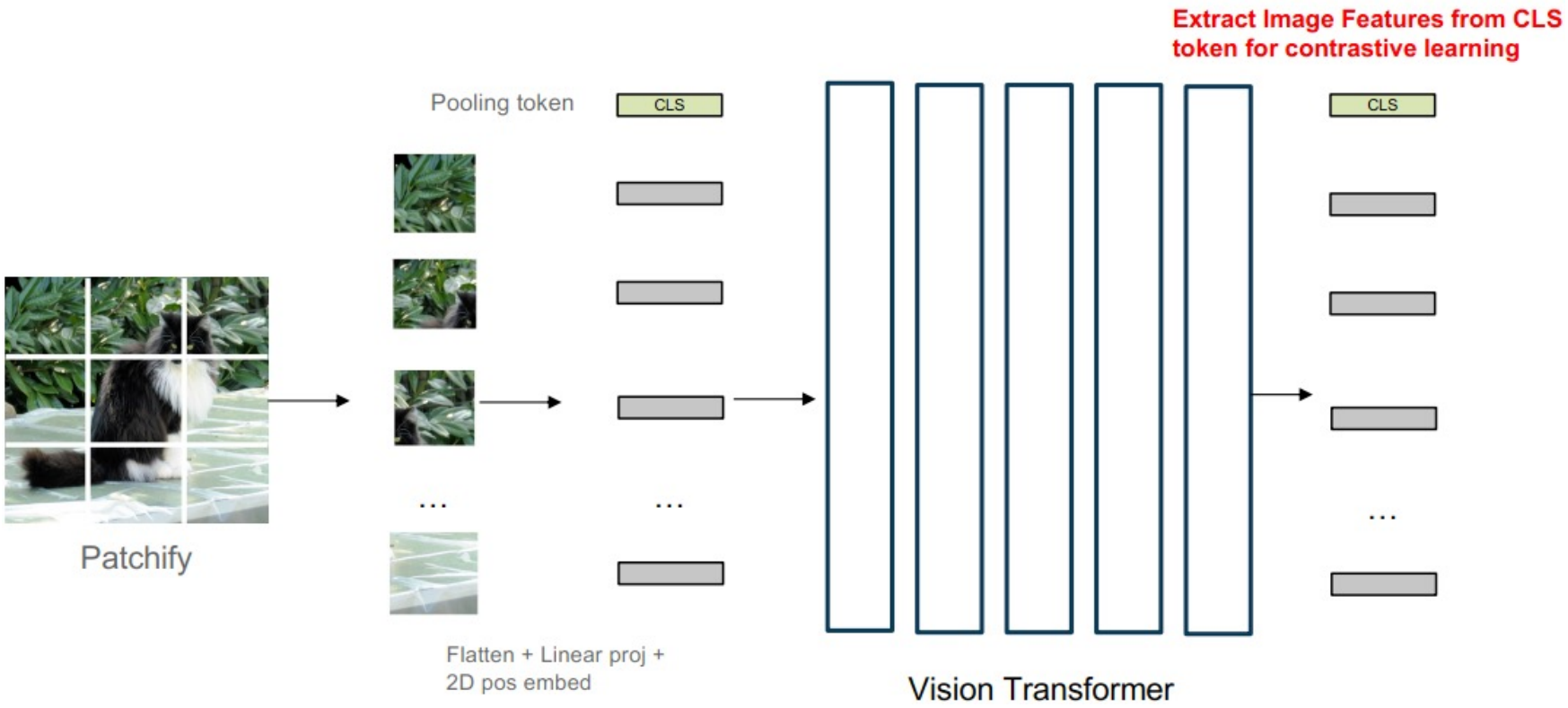
# LLaVA: The CLIP encoder is a good option!

At the end of training, you have a model that will give you a similarity score between an image and a text




https://cs231n.stanford.edu/

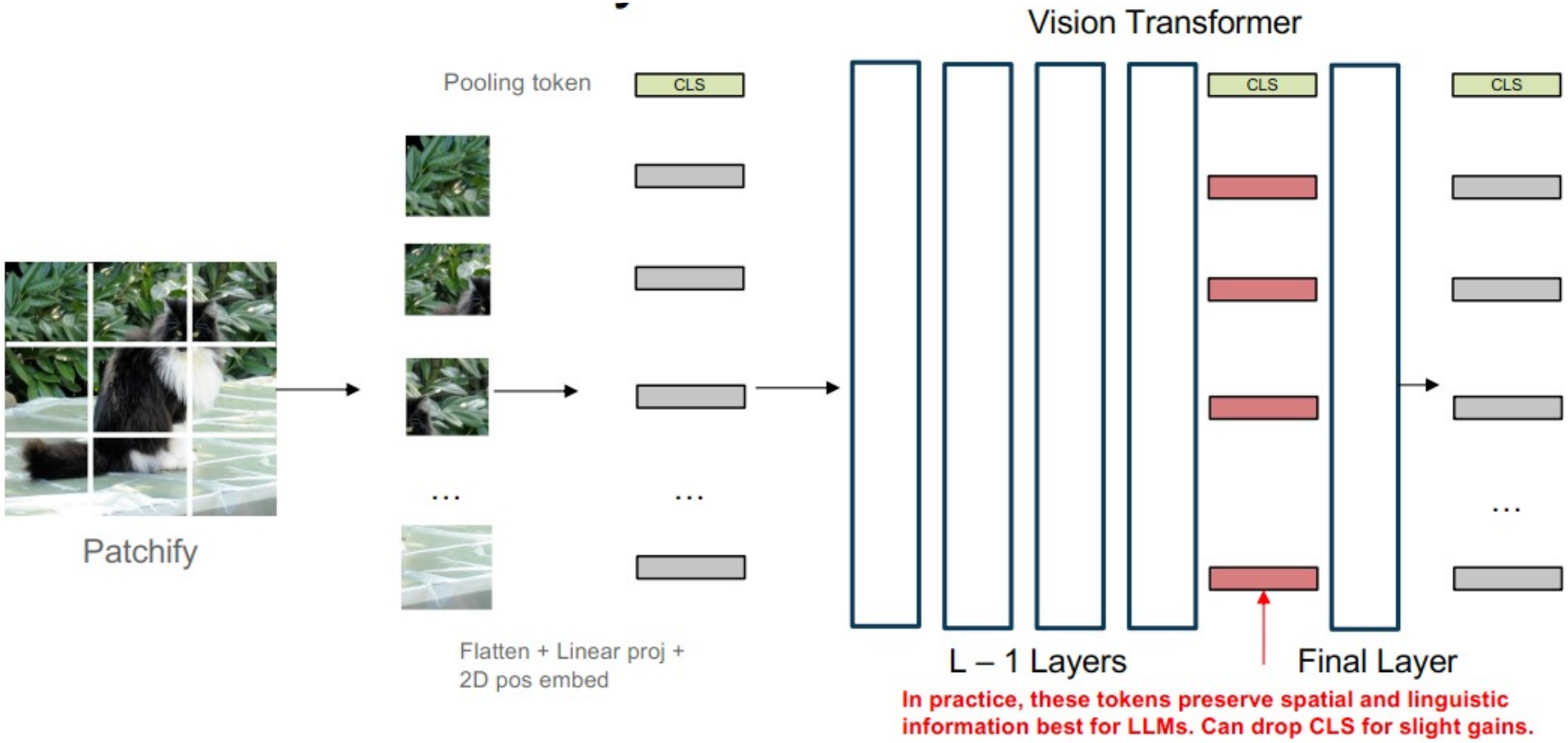
# LLaVA: What features should we use from CLIP?




https://cs231n.stanford.edu/

Image Source: https://en.wikipedia.org/wiki/File:Cat_outside.jpg

# LLaVA: Use Penultimate Layer!




https://cs231n.stanford.edu/

Image Source: https://en.wikipedia.org/wiki/File:Cat_outside.jpg

# LLaVA: Overall Architecture + Training Recipe

1. Initialize with pretrained Language Model for LLM Decoder (e.g. LLaMA) and pretrained image encoder (e.g. CLIP)
2. Train a new linear layer to bridge CLIP features to LLM input space
3. Finetune LLM + linear layer together

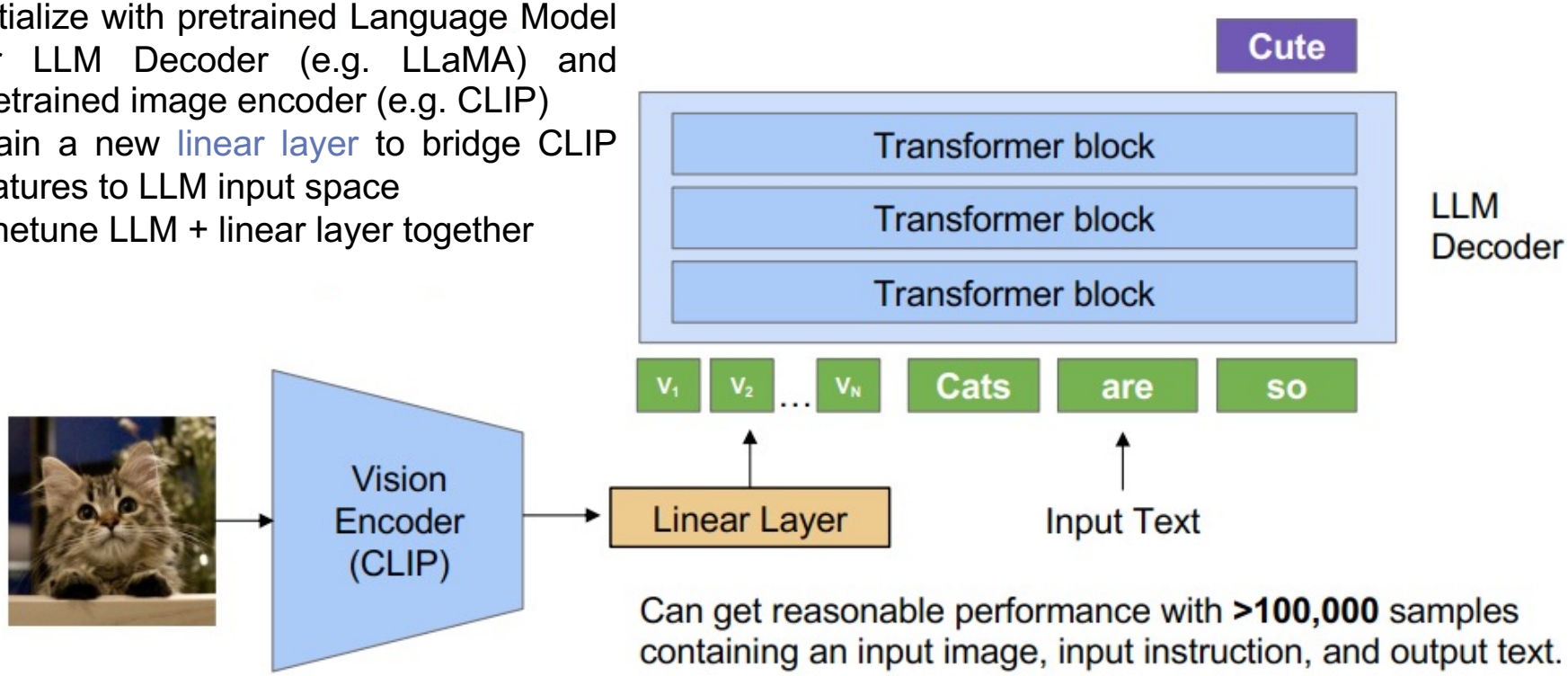
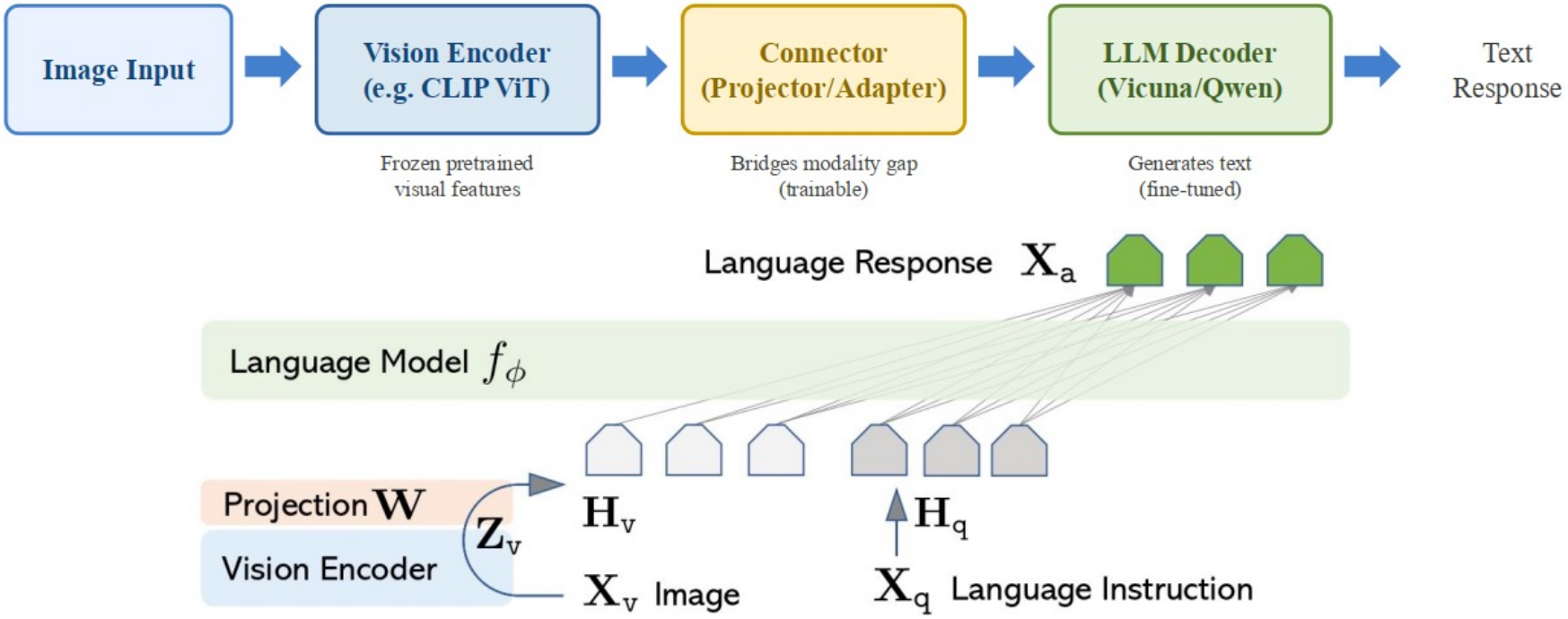




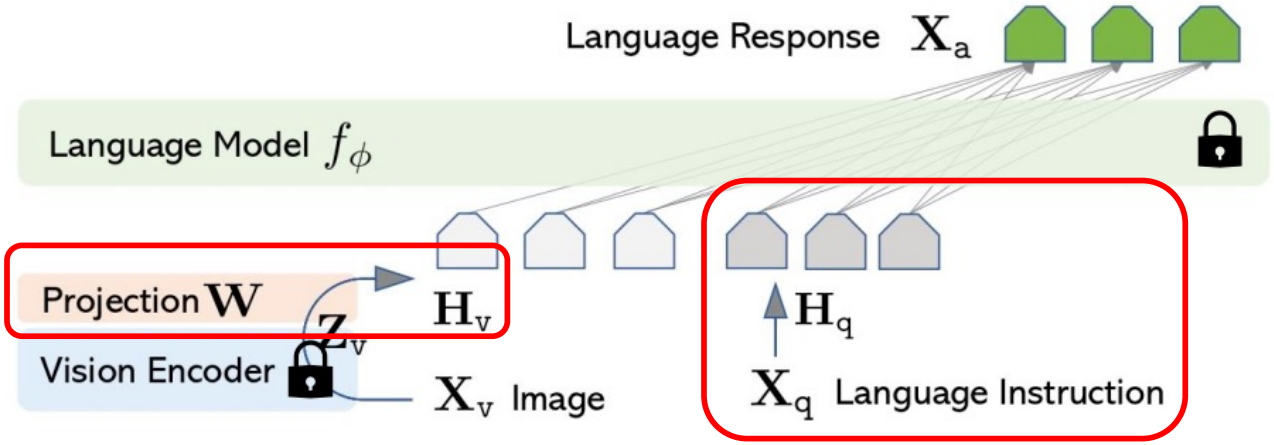

Image Source: https://en.wikipedia.org/wiki/File:Cat_outside.jpg

# Three Core Components of Llava-style Vision-Language Models




Source: [Liu et al. 2023]

# Three Core Components of Llava-style Vision-Language Models




- Training Stage 1 / 2
  - large-scale image-text pretraining with **the LLM and Visual Encoder Frozen**
  - Data: 600K filtered CC3M

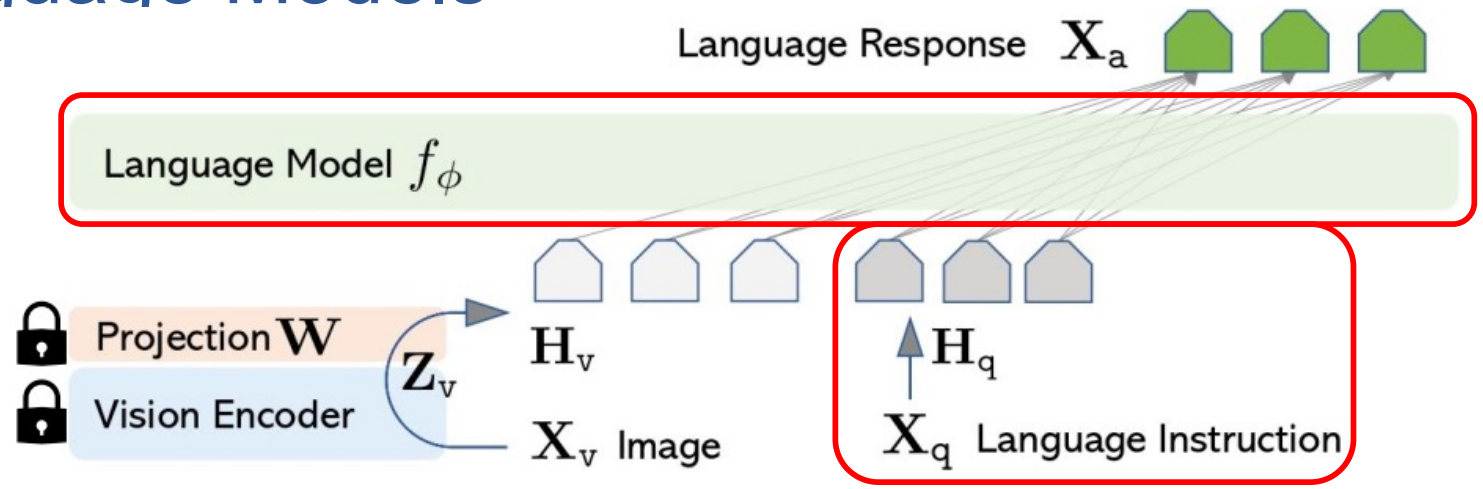


```
a river has burst
it 's banks and has
spread out onto
arable farmland
alongside
```

Source: [Liu et al. 2023]

https://www.cs.rice.edu/~vo9/deep-vislang/

# Three Core Components of Llava-style Vision-Language Models

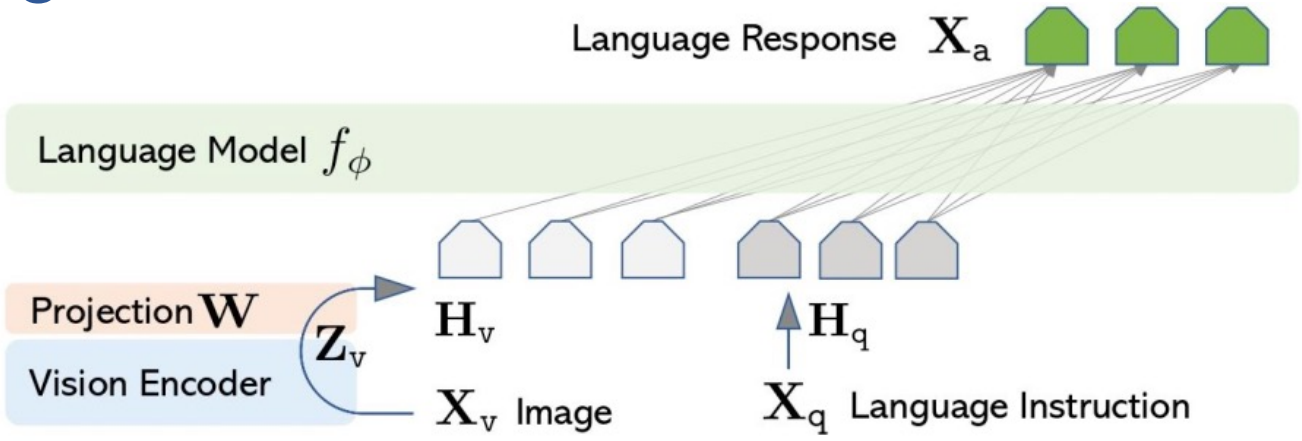



- Training Stage 2 / 2
  - visual instruction tuning stage: use 150K GPT-generated multimodal instruction-following data

Source: [Liu et al. 2023]

https://www.cs.rice.edu/~vo9/deep-vislang/

# Three Core Components of Llava-style Vision-Language Models

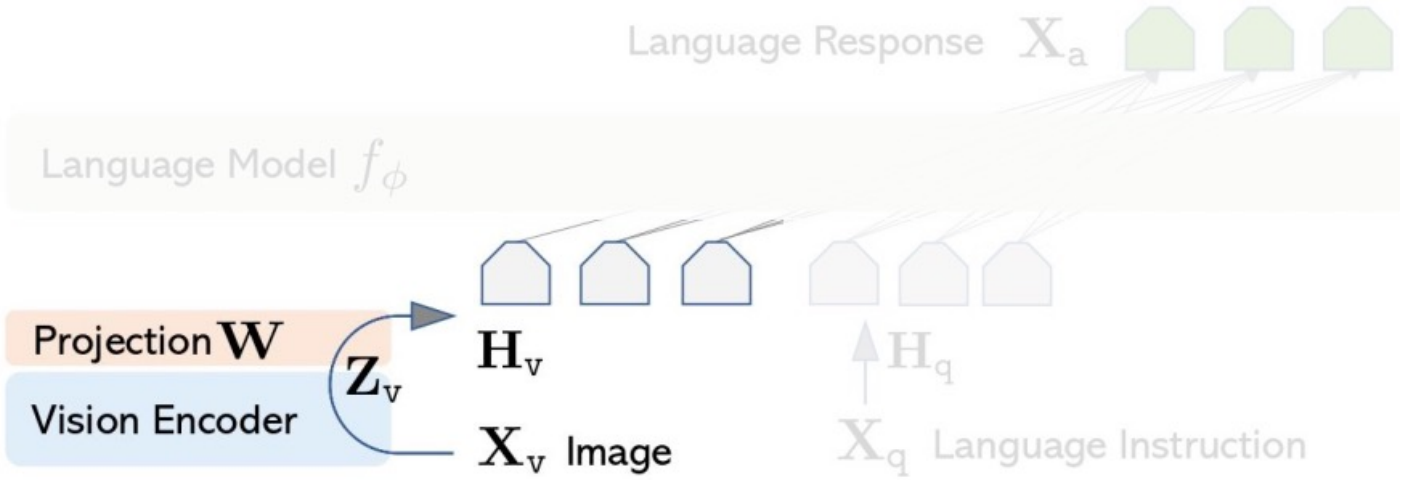



- LLaVA-1.5 Update
  - Projection Layer: Linear Projection -> MLP
  - Visual Encoder: CLIP ViT-L/14 -> **CLIP-ViT-L-336px**
  - Data Mix: Stage 1 - **558K LAION-CC-SBU**; Stage 2: 150K + **515K VQA data from academic-oriented tasks**

Source: [Liu et al. 2023]

https://www.cs.rice.edu/~vo9/deep-vislang/

# The Bottleneck of LLaVA: Fixed Visual Tokens for Every Image

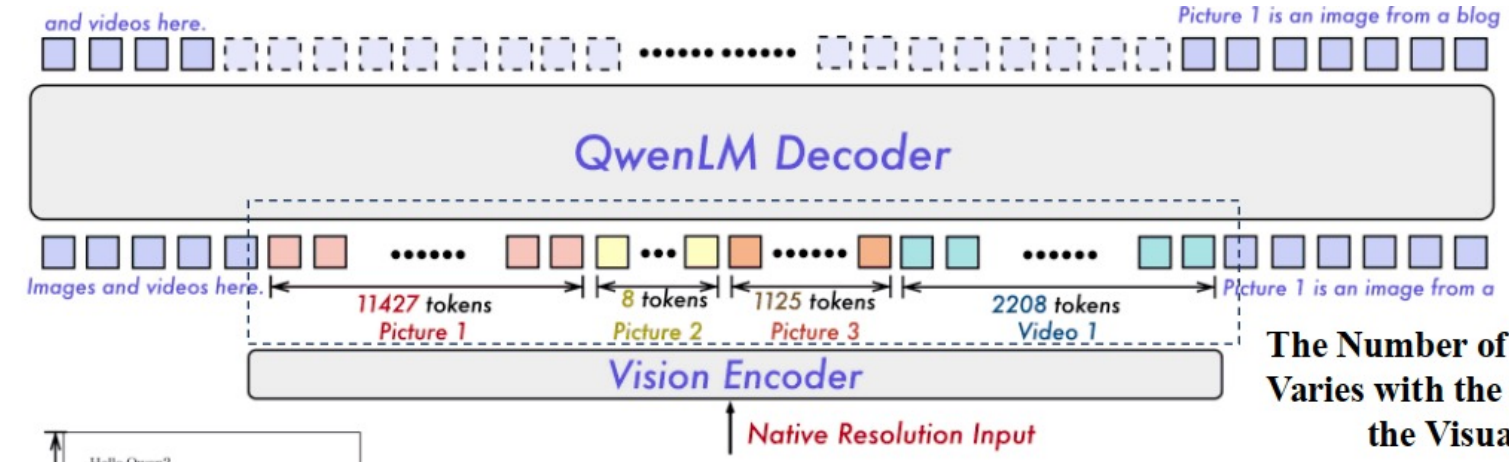



- **The limitation of early VLMs (LLaVA 1.5 and before)**
  - **Each image gets resized to 224x224 and processed by a CLIP ViT.**
  - **224x224 with patch size:14 ->256 patches per image**
  - **Details are not preserved at all.**

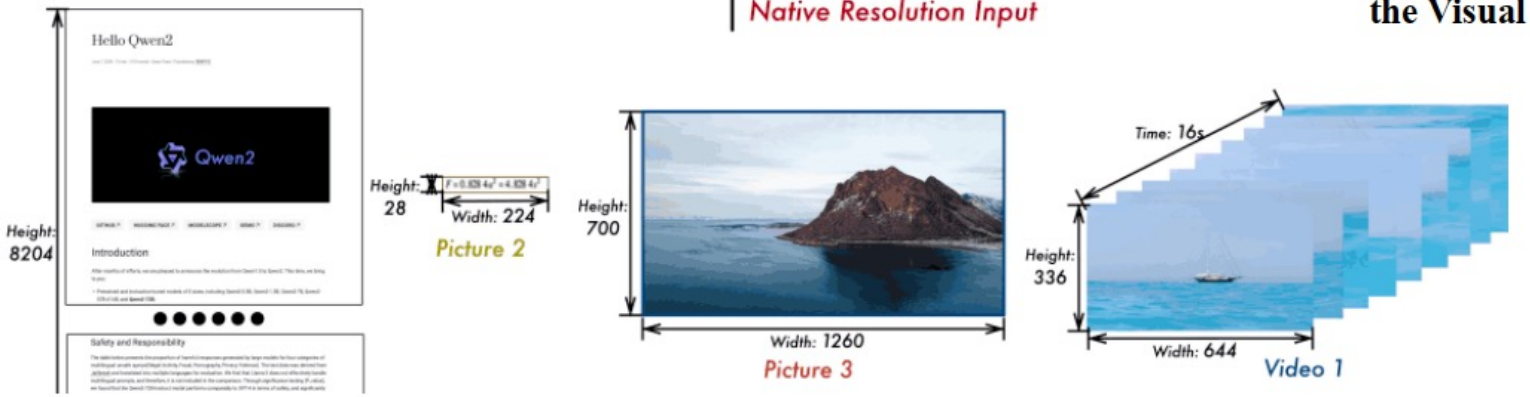
Source: [Liu et al. 2023]

https://www.cs.rice.edu/~vo9/deep-vislang/

# Solution: Vision-Language Models with Dynamic Resolution




**The Number of Visual Tokens Varies with the Complexity of the Visual Input!**




https://www.cs.rice.edu/~vo9/deep-vislang/

https://arxiv.org/pdf/2407.10671

# Solution: Vision-Language Models with Dynamic Resolution

main Qwen2-VL-7B-Instruct / preprocessor_config.json

yangapku Initial commit f330ecb

raw Copy download link history blame contribute delete

```
{
  "min_pixels": 3136,
  "max_pixels": 12845056,
  "patch_size": 14,
  "temporal_patch_size": 2,
  "merge_size": 2,
  "image_mean": [
    0.48145466,
    0.4578275,
    0.40821073
  ],
  "image_std": [
    0.26862954,
    0.26130258,
    0.27577711
  ],
  "image_processor_type": "Qwen2VLImageProcessor",
  "processor_class": "Qwen2VLProcessor"
}
```

- **Step 1: Resize but keep aspect ratio**
  - If an image is too small, it can be upscaled to exceed **min_pixels.**
  - If it's too large, it can be downscaled to stay under **max_pixels.**





Resize



Resize w/ aspect ratio

https://www.cs.rice.edu/~vo9/deep-vislang/

https://arxiv.org/pdf/2407.10671

# Solution: Vision-Language Models with Dynamic Resolution

main ▾ Qwen2-VL-7B-Instruct / preprocessor_config.json

yangapku Initial commit f330ecb

raw | Copy download link | history | blame | contribute | delete

```
{
  "min_pixels": 3136,
  "max_pixels": 12845056,
  "patch_size": 14,
  "temporal_patch_size": 2,
  "merge_size": 2,
  "image_mean": [
    0.48145466,
    0.4578275,
    0.40821073
  ],
  "image_std": [
    0.26862954,
    0.26130258,
    0.27577711
  ],
  "image_processor_type": "Qwen2VLImageProcessor",
  "processor_class": "Qwen2VLProcessor"
}
```

➢ **Step 2: Patchify**

- If resized to 3584×3584: 3584/14 = 256 → 256×256 = 65,536 patch tokens
- If resized to 56×56: 56/14 = 4 → 4×4 = 16 patch tokens



**The number of patches is no longer a fixed number.**

https://www.cs.rice.edu/~vo9/deep-vislang/

https://arxiv.org/pdf/2407.10671

# Solution: Vision-Language Models with Dynamic Resolution

Modified ViT to use 2D-RoPE positional embeddings

main Qwen2-VL-7B-Instruct / preprocessor_config.json

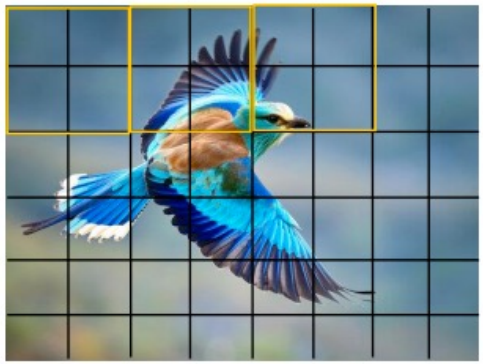
yangapku Initial commit f330ecb

raw Copy download link history blame contribute delete

```
{
  "min_pixels": 3136,
  "max_pixels": 12845056,
  "patch_size": 14,
  "temporal_patch_size": 2,
  "merge_size": 2,
  "image_mean": [
    0.48145466,
    0.4578275,
    0.40821073
  ],
  "image_std": [
    0.26862954,
    0.26130258,
    0.27577711
  ],
  "image_processor_type": "Qwen2VLImageProcessor",
  "processor_class": "Qwen2VLProcessor"
}
```

## ➢ Step 3: Token compression

- Qwen2-VL applies a simple MLP "compression" that merges **adjacent tokens** in token space so the LLM doesn't get overwhelmed by huge grids.
- Indicated by "merge_size" in the config.



if 48 patch tokens -> 12 visual tokens in the LLM input

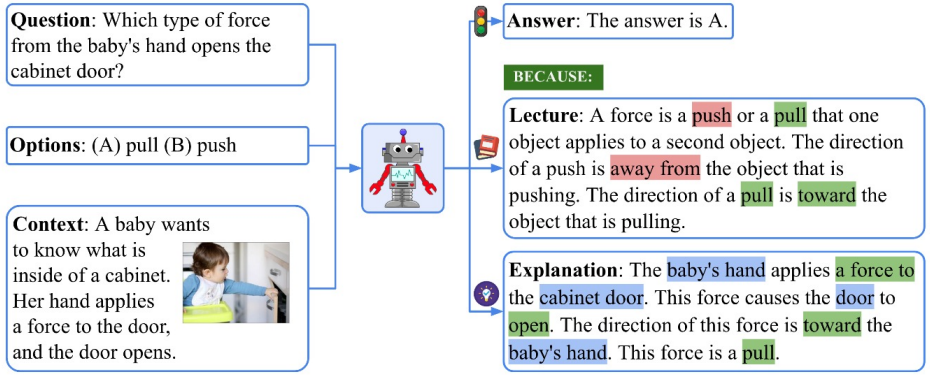
https://www.cs.rice.edu/~vo9/deep-vislang/

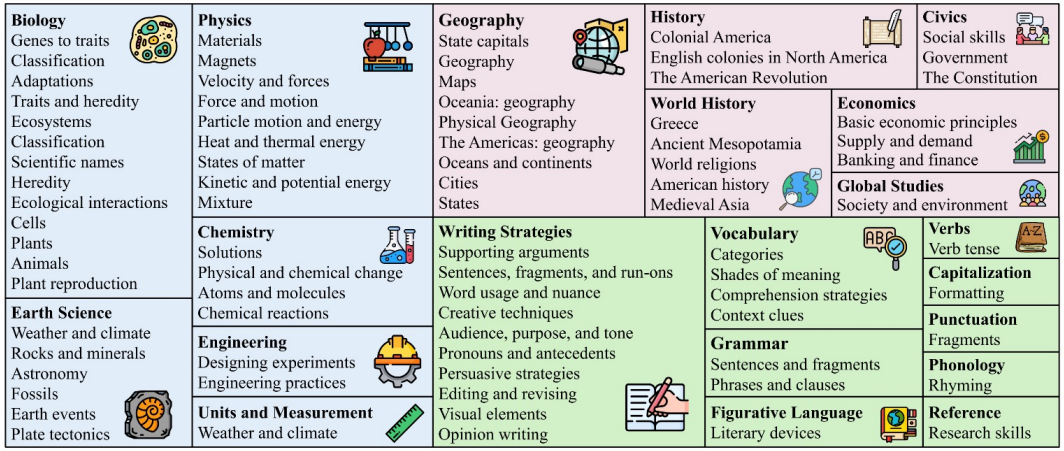
https://arxiv.org/pdf/2407.10671

# MLLM Functionality: Image Perception - LLaVA

Evaluation: ScienceQA (Lu et al., NeurIPS 2022)

- 21k multimodal multiple-choice questions

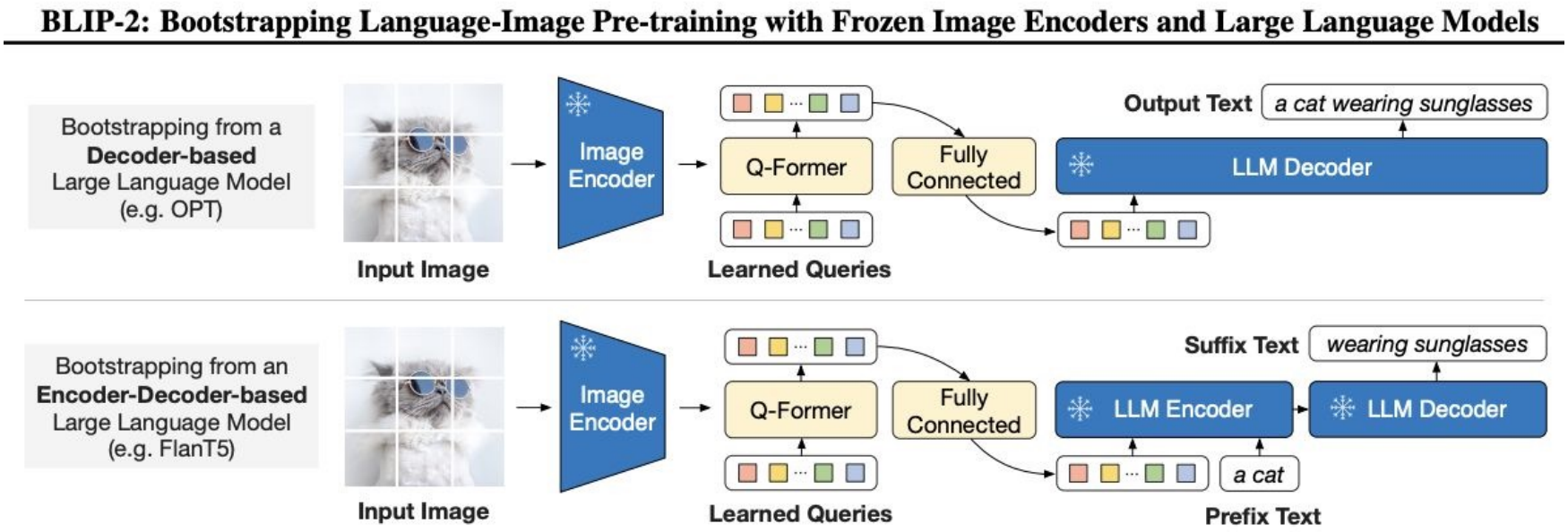


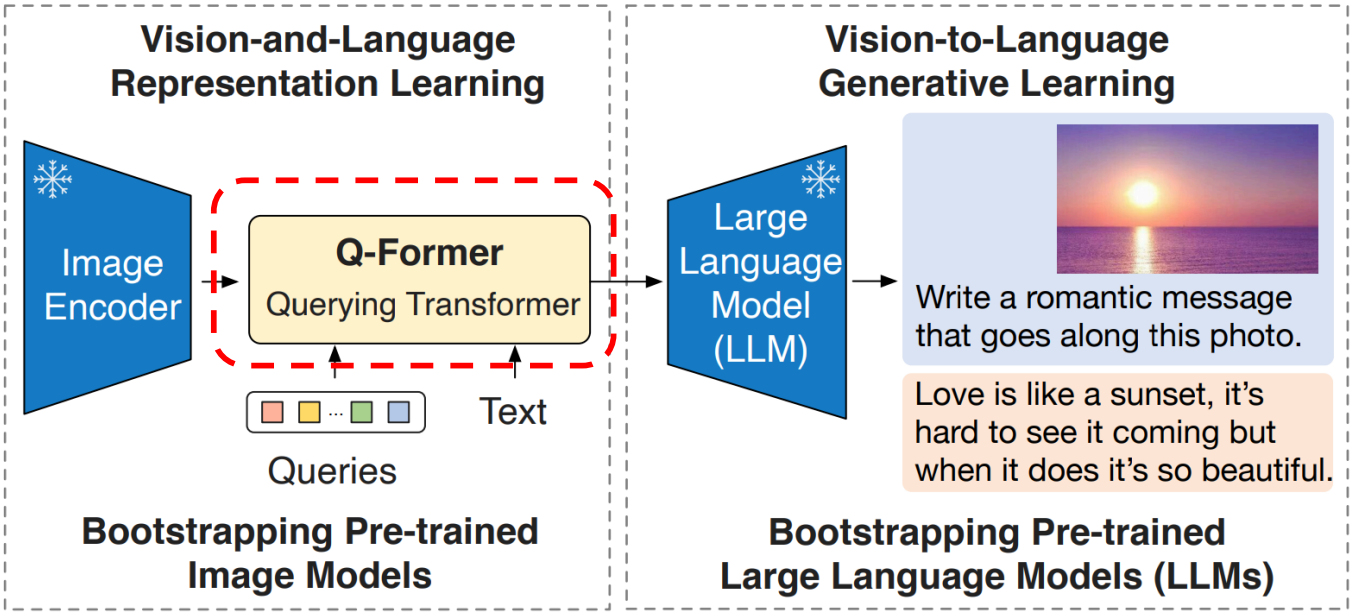

Kyle Buettner

# MLLM Functionality: Image Perception - BLIP/BLIP2 (Li et al., 2023)

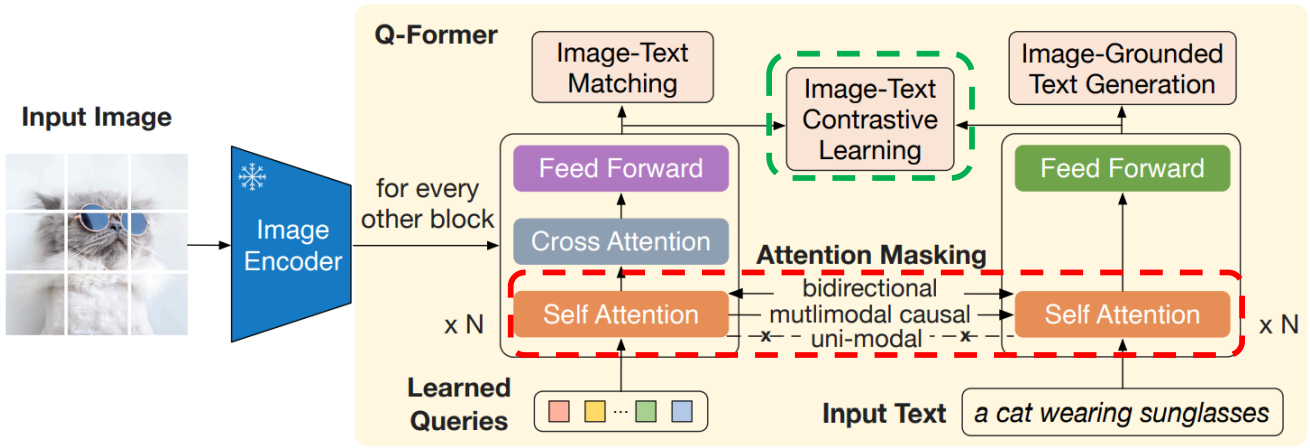
Freeze it all (CLIP-ViT / OPT decoder / FlanT5 encoder-decoder)



# BLIP2: Bootstrapping language-image pre-training with frozen image encoders and large language models



# BLIP2 Architecture: Q-Former

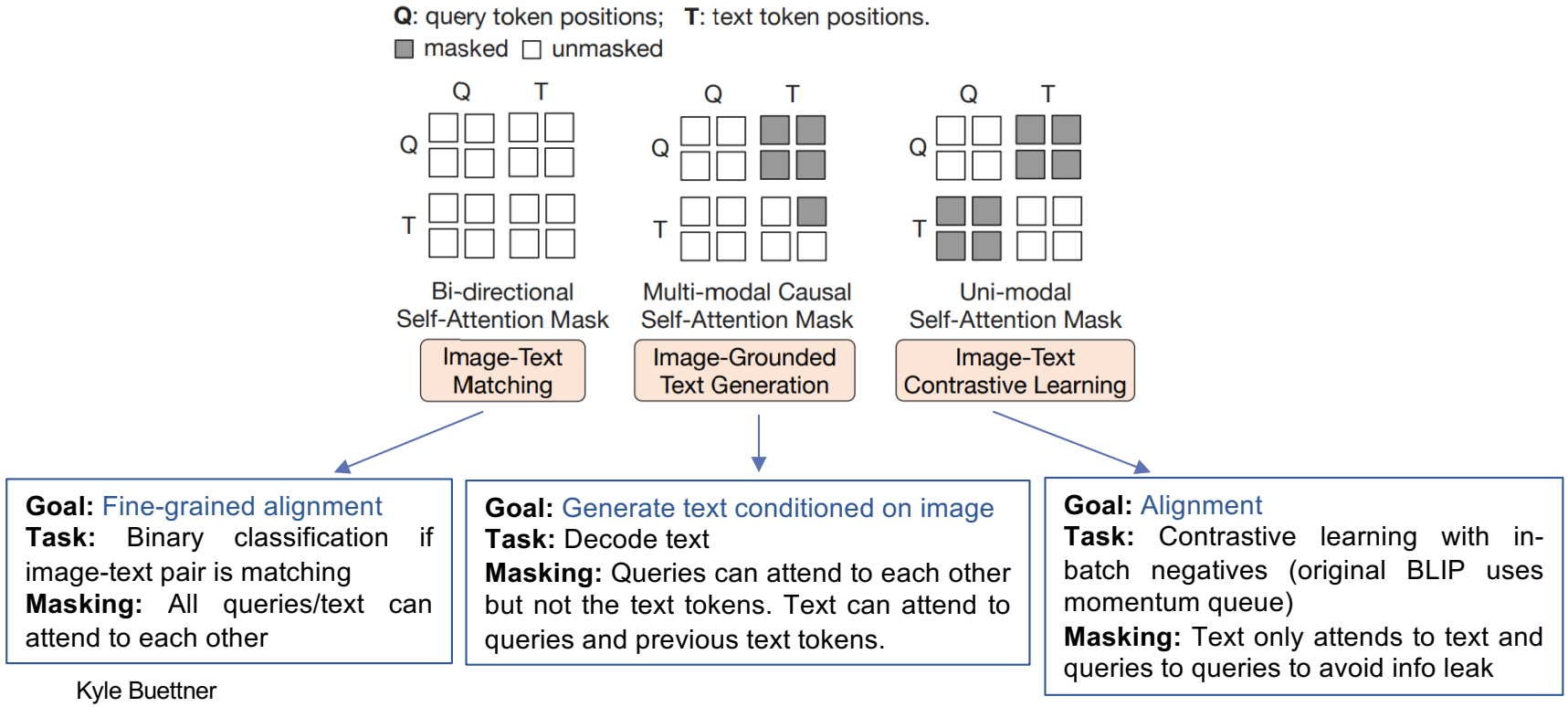



- Extracts fixed # of features from image encoder
- Has image and text transformers with same self-attention layers
- Learnable query embeddings ***(Z)*** are inputs to image transformer
  - 32x768; can interact with each other, text, and frozen image features
  - **Goal:** Extract visual info most relevant to the text
- Initialized with $BERT_{base}$ weights
- 188M parameters

Kyle Buettner

# BLIP2 - Stage 1 Training: Representation Learning

- Three objectives are jointly optimized using different self-attention masking strategies to control query-text interaction




Kyle Buettner

# BLIP2 - Stage 2: Generative Learning

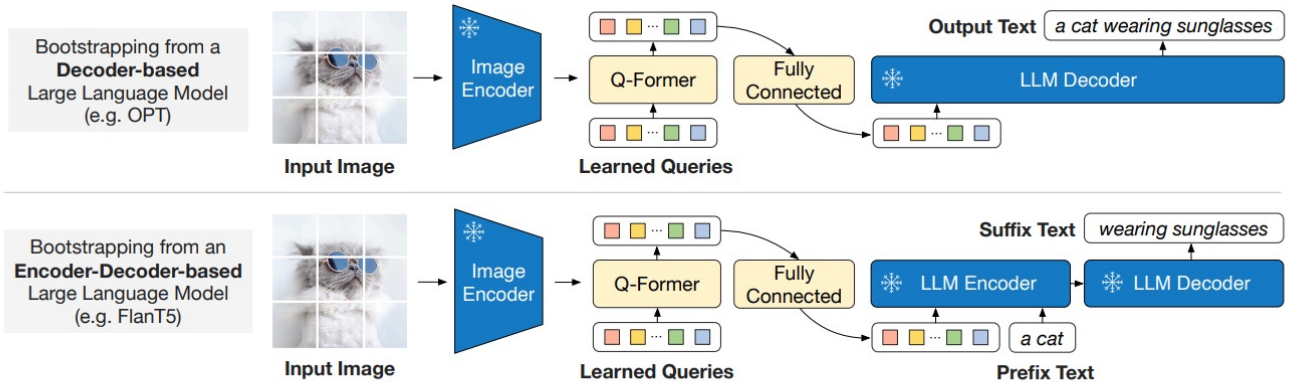
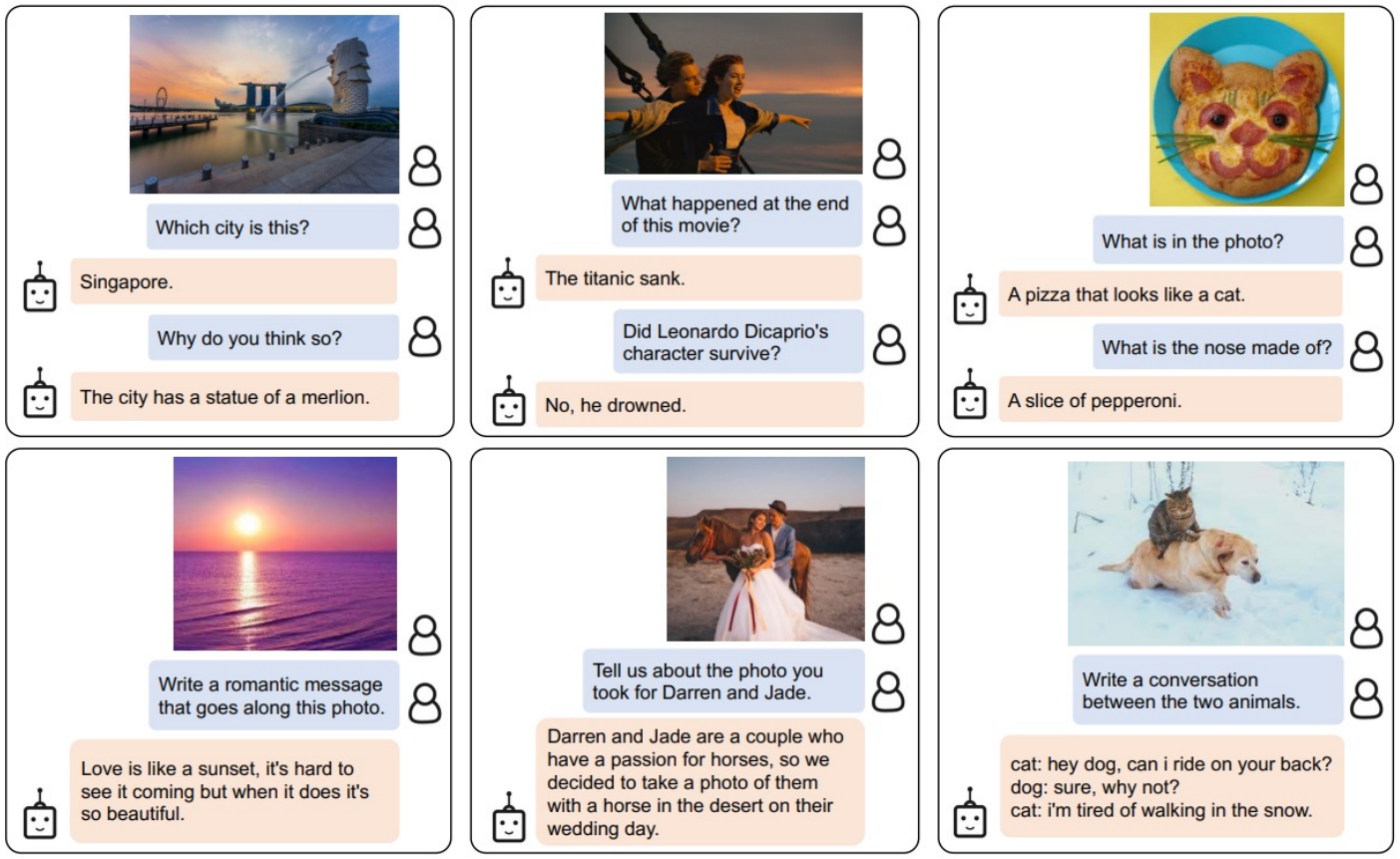



*Figure 3.* BLIP-2's second-stage vision-to-language generative pre-training, which bootstraps from frozen large language models (LLMs). (**Top**) Bootstrapping a decoder-based LLM (e.g. OPT). (**Bottom**) Bootstrapping an encoder-decoder-based LLM (e.g. FlanT5). The fully-connected layer adapts from the output dimension of the Q-Former to the input dimension of the chosen LLM.

- Query embeddings ***Z*** projected into LLM embedding space and prepended to input text embeddings
  - A soft visual prompting mechanism for the LLM
- Different loss forms are used for different LM architectures
  - Decoder – language modeling loss over the queries
  - Encoder-decoder – prefix pretrained with LM loss; suffix used as generation target

# BLIP2 - Instructed Zero-Shot Image-to-Text Generation (2/2)




Kyle Buettner

# BLIP2 - Example Issues

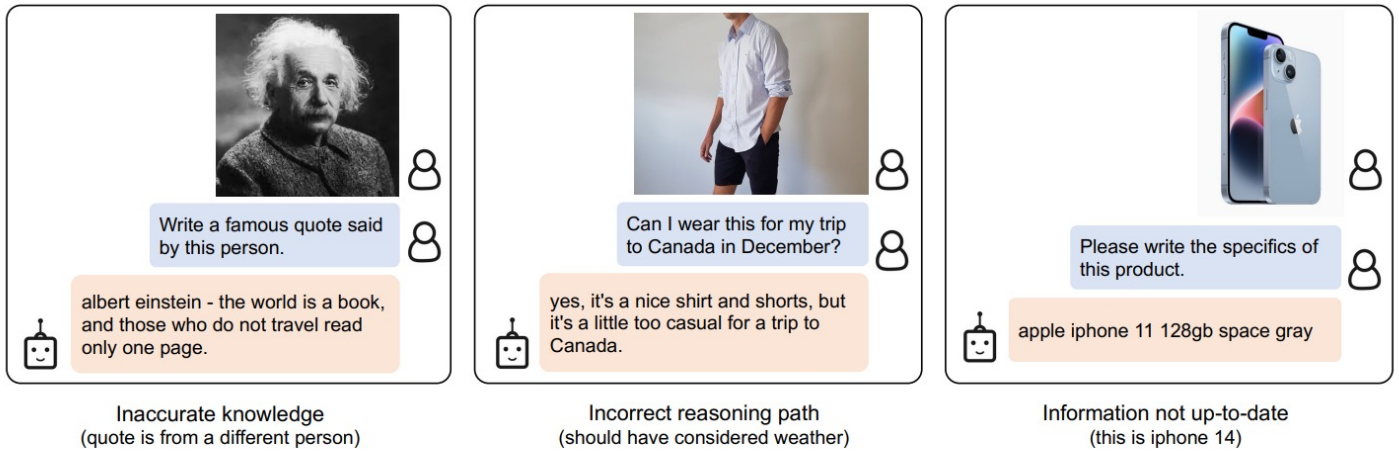



*Figure 6.* Incorrect output examples for instructed zero-shot image-to-text generation using a BLIP-2 model w/ ViT-g and FlanT5$_{XXL}$.

# BLIP2 – Expansion --> Meta-queries

In this paper, we aim to deliver the promise of unified models via a simpler philosophy: *Render unto diffusion what is generative, and unto LLMs what is understanding.*

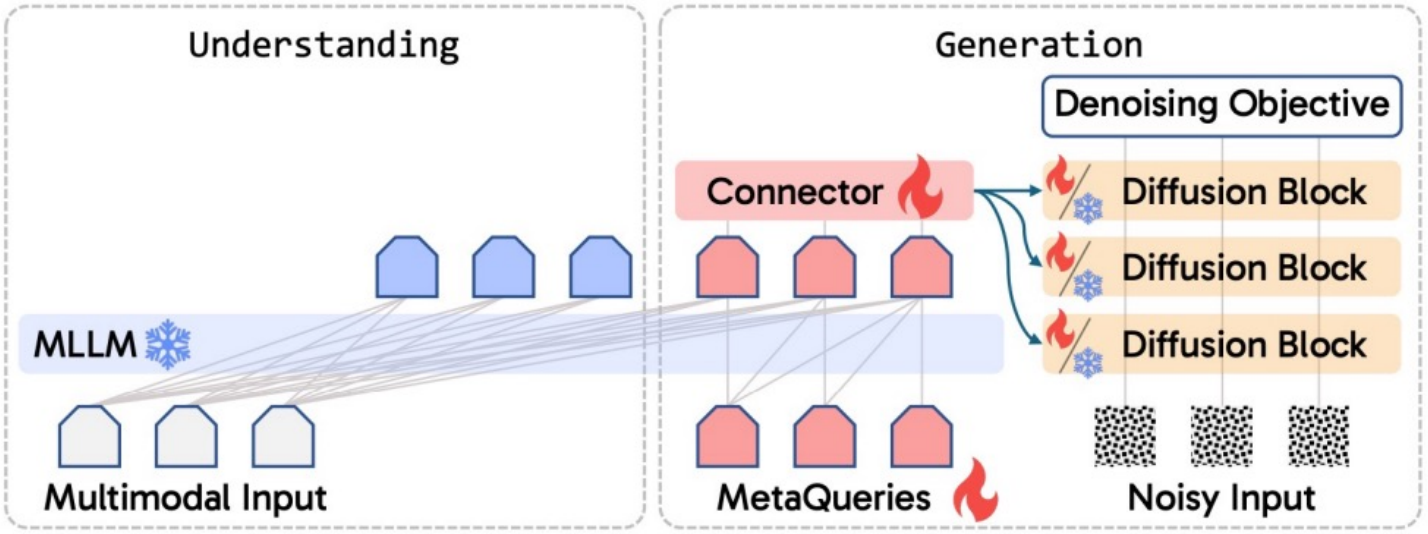
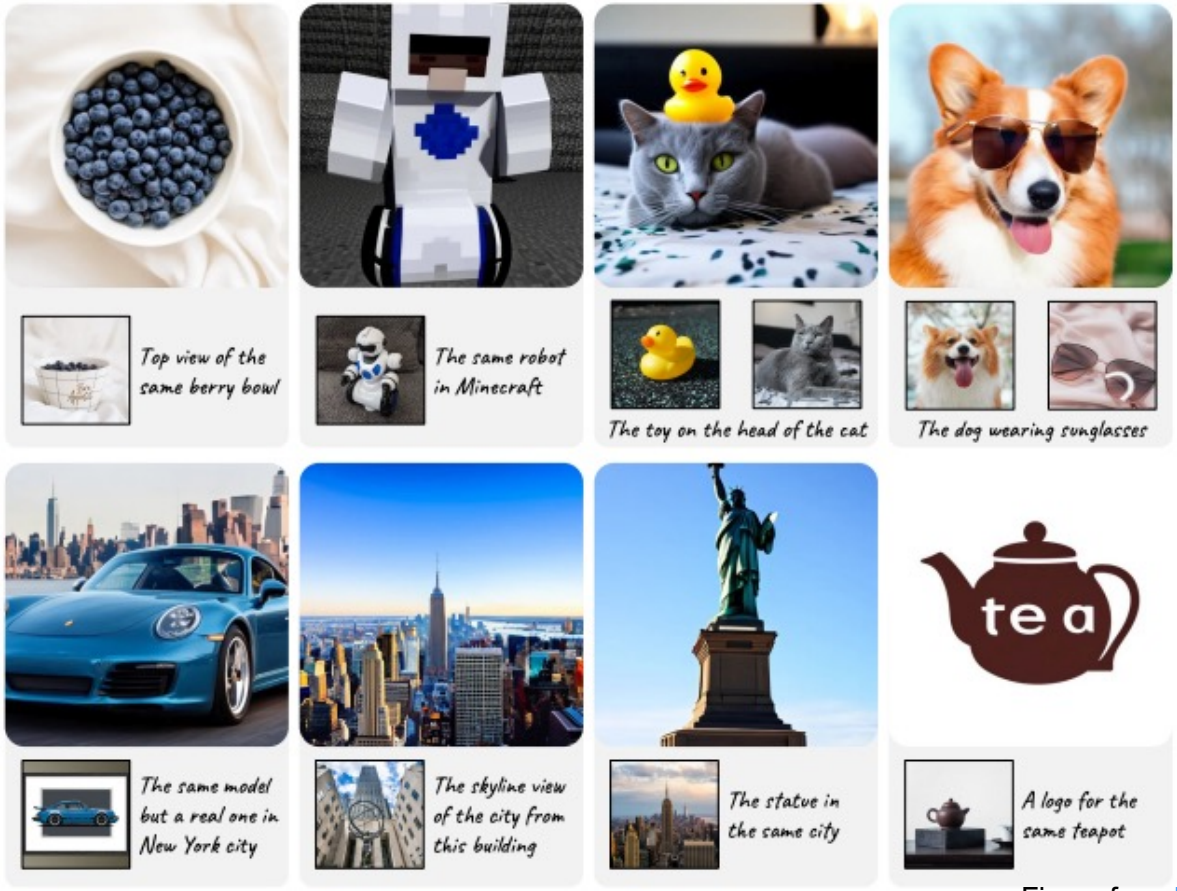



**Figure 1** Overview of our model. Blue tokens maintain SOTA multimodal understanding; `MetaQueries` are learnable queries that directly applied to frozen MLLMs to query out conditions for generation. The model is tuned using only denoising objective with paired data. The generative diffusion models can be either frozen or further instruction-tuned for advanced generation tasks.

https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://arxiv.org/pdf/2504.06256

# BLIP2 – Expansion --> Meta-queries

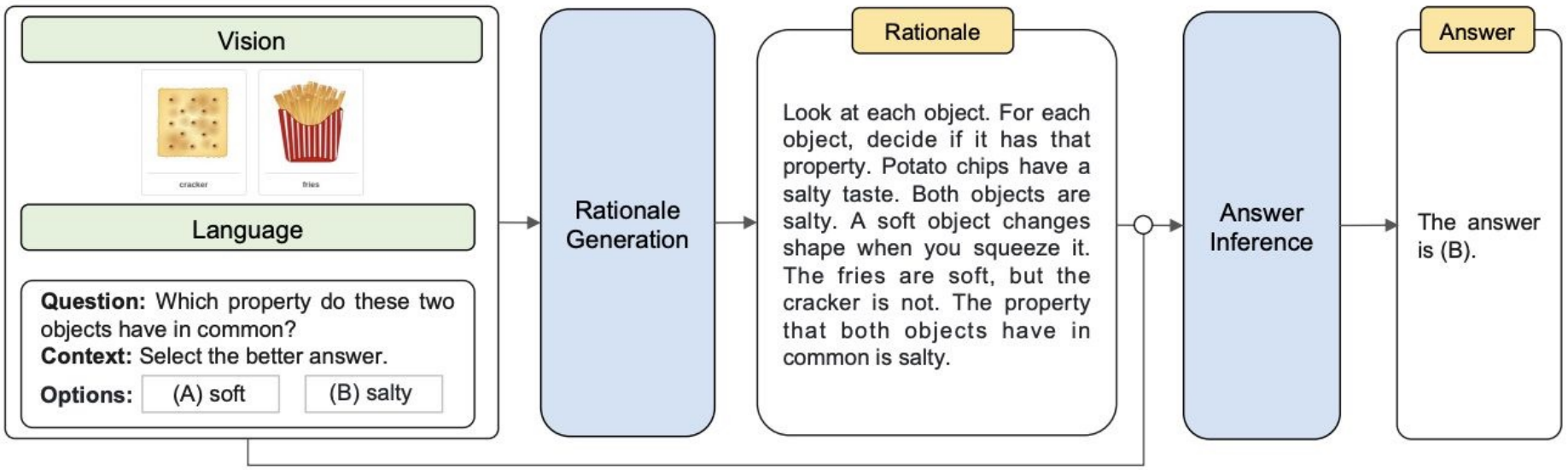



https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://arxiv.org/pdf/2504.06256

# MLLM Functionality: Image Perception - Multimodal "Chain of Thought" (Zhang et al., 2023)

Providing a rationale helps give the right answer.




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception - KOSMOS-1 (Huang et al., 2023)

LLMs => MLLMs == Foundation Model

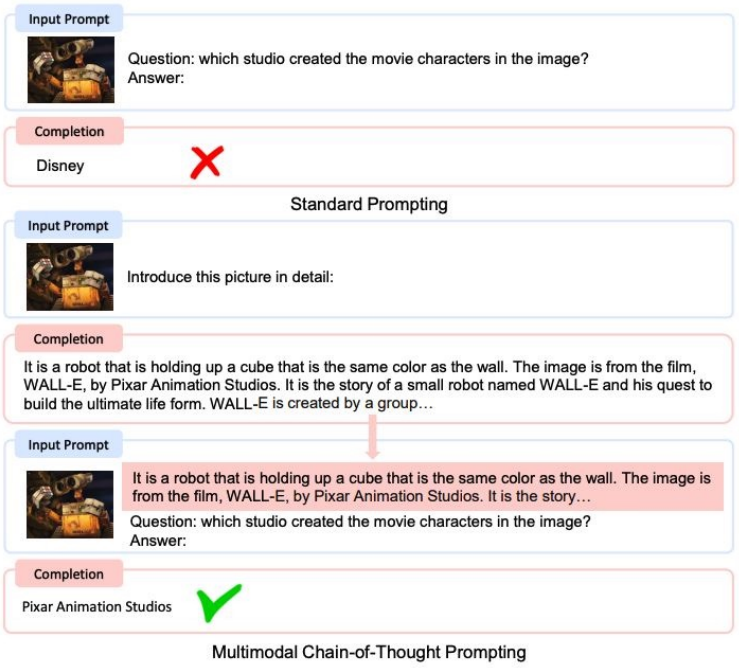
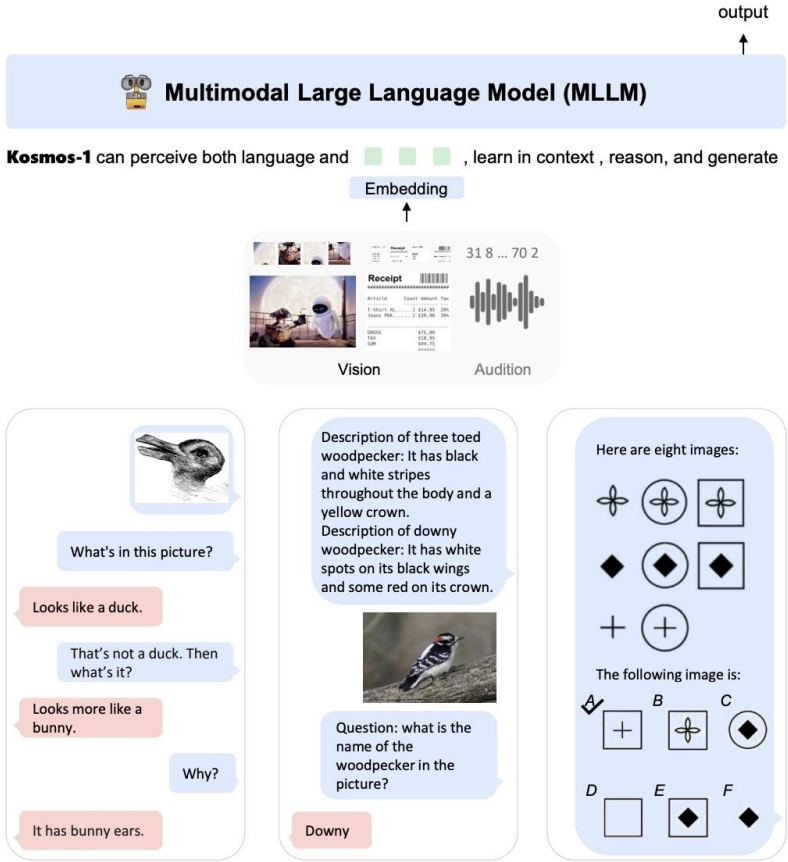
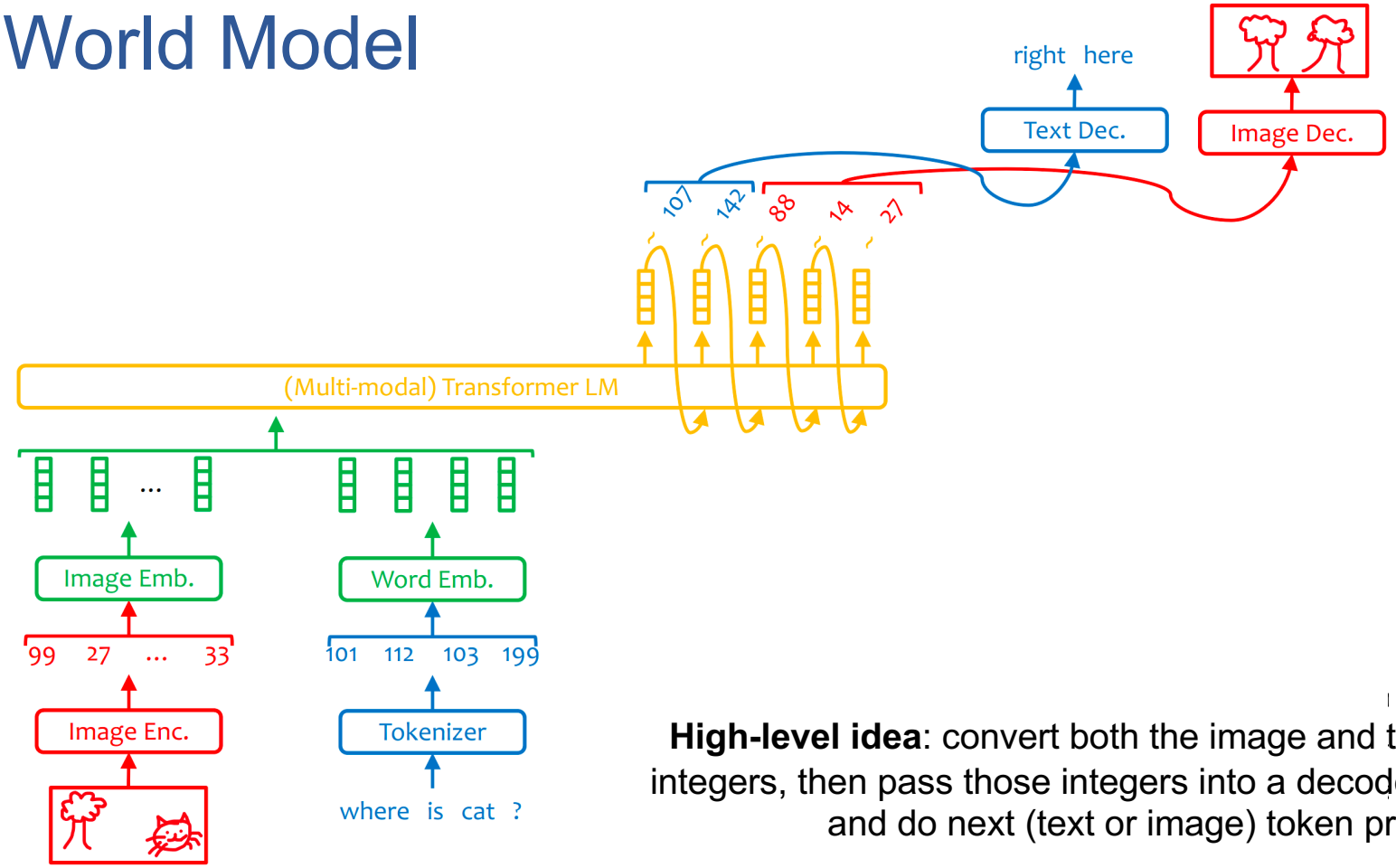



Figure 5: Multimodal Chain-of-Thought prompting enables KOSMOS-1 to generate a rationale first, then to tackle complex question-answering and reasoning tasks.




Stanford CS224N - *Douwe Kiela* - Multimodal Deep Learning – https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/

# MLLM Functionality: Image Perception – Large World Model



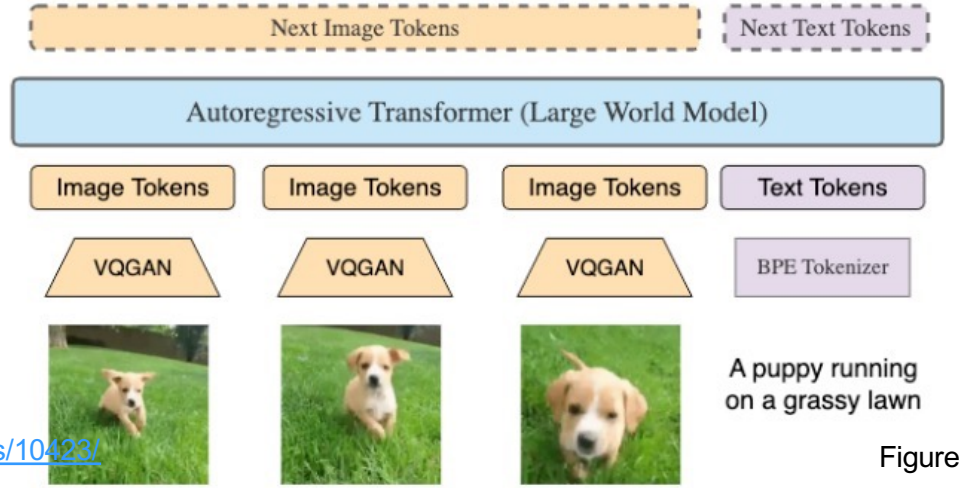

**High-level idea**: convert both the image and the text inputs into integers, then pass those integers into a decoder-only transformer and do next (text or image) token prediction

# MLLM Functionality: Image Perception – Large World Model

The **Large World Model (LWM)** is an example of a Transformer LM that works with both discrete text tokens and discrete image tokens

**Key idea**:

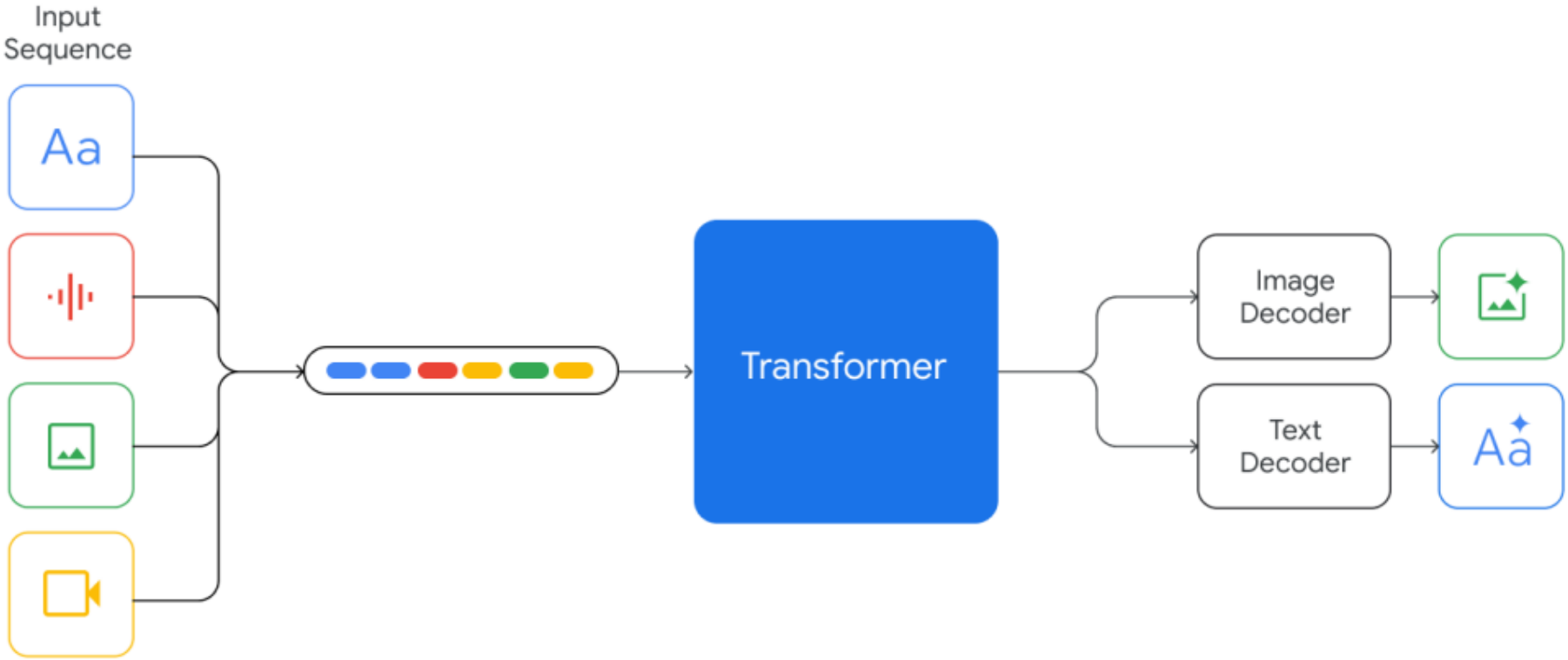
- pretrain your image tokenizer/de-tokenizer (VQGAN)
- any images in your data can be converted to their discrete representations ahead of time
- then train your Transformer LM as you would any other LM on discrete tokens
- at test time, whenever you see a sequence of image tokens, convert them back to an image




https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from http://arxiv.org/abs/2402.08268

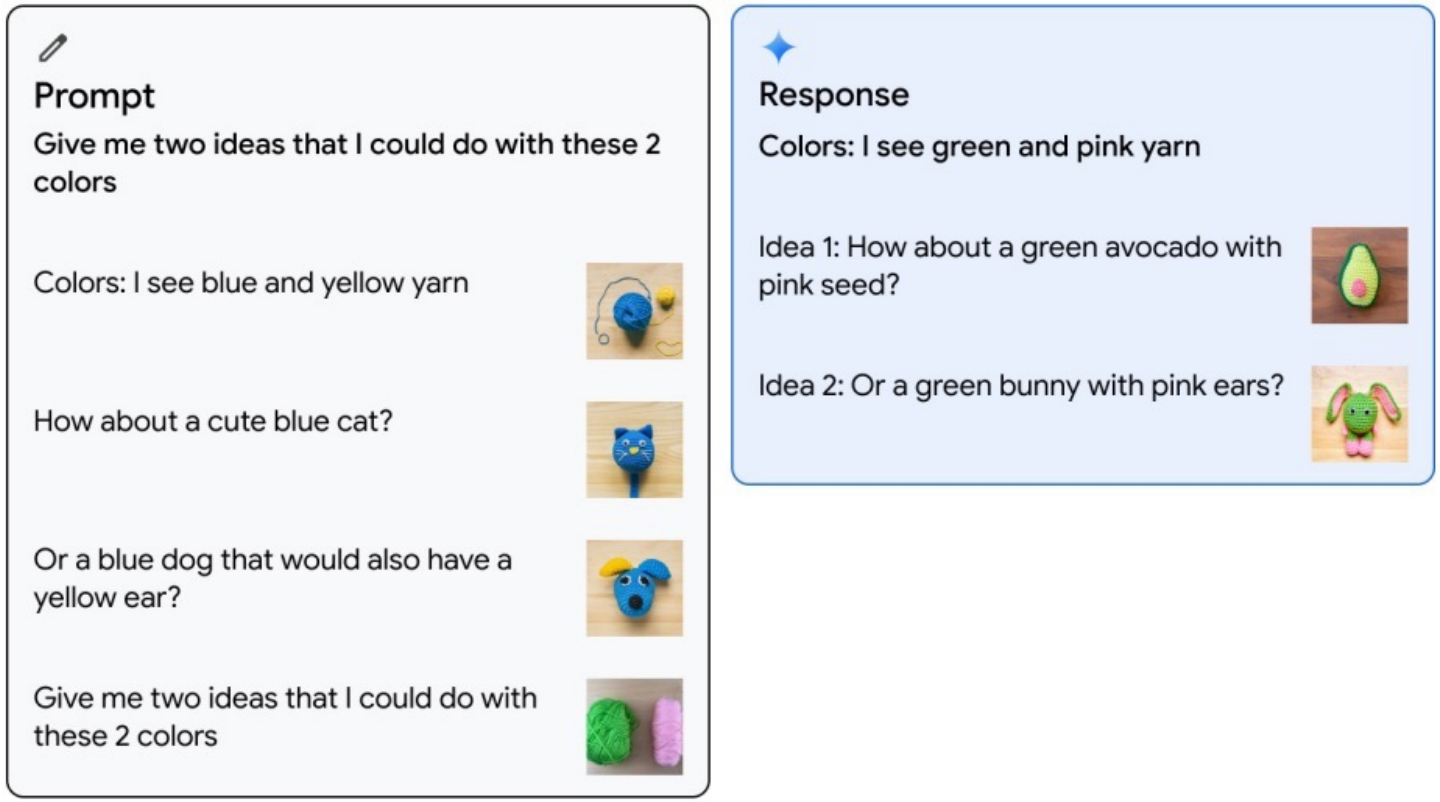
# MLLM Functionality: Image Perception – Gemini




https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://arxiv.org/pdf/2312.11805

# MLLM Functionality: Image Perception – Gemini




https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://arxiv.org/pdf/2312.11805

To join, go to: **ahaslides.com/FJWAK**

AhaSlides

# Use Gemini, to generate and submit creative visual ideas (images). Search for inspiration from slide 91 from Multimodal Language Models.

Get Feedback

Edit slides

Add slides

Group into themes

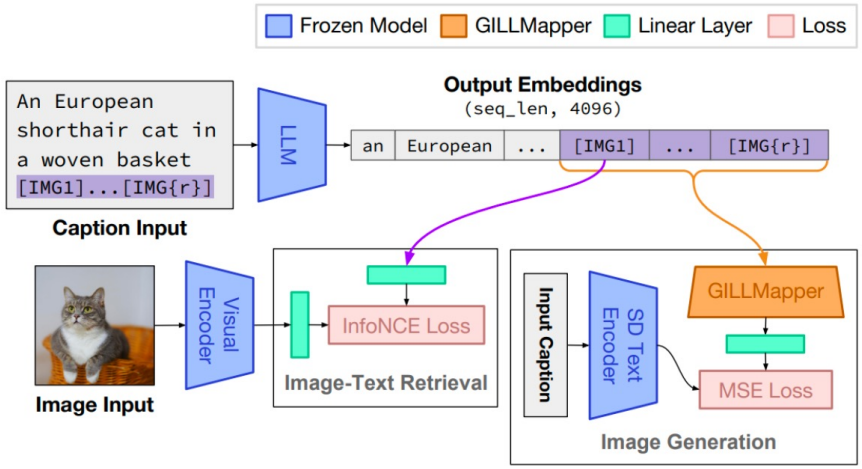
0 0 /100

# MLLM Functionality: Generation

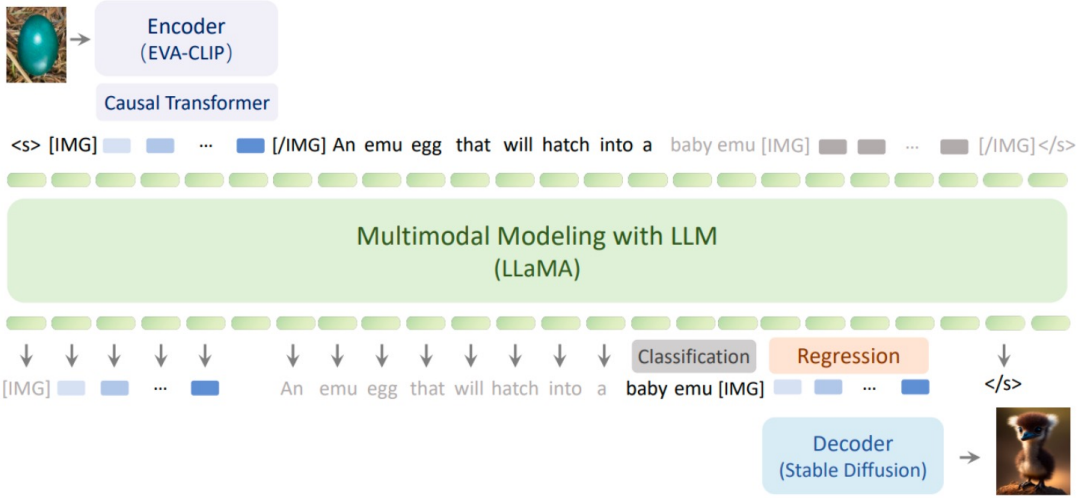
| | Modality (w/ Language) | | | |
|---|---|---|---|---|
| | **Image** | **Video** | **Audio** | **3D** |
| Input-side Perceiving | Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, … | VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, … | AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, … | 3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, … |
| | [Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, … | [Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, … | - | - |
| | Video-LLaVA, Chat-UniVi, LLaMA-VID | | - | - |
| | Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, … | | | - |
| Perceiving + Generating | GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, … | GPT4Video, Video-LaVIT, VideoPoet, … | AudioGPT, SpeechGPT, VIOLA, AudioPaLM, … | - |
| | [Pixel-wise] Vitron | | - | - |
| | NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, … | | | - |

# MLLM Functionality: Image Generation

## GILL



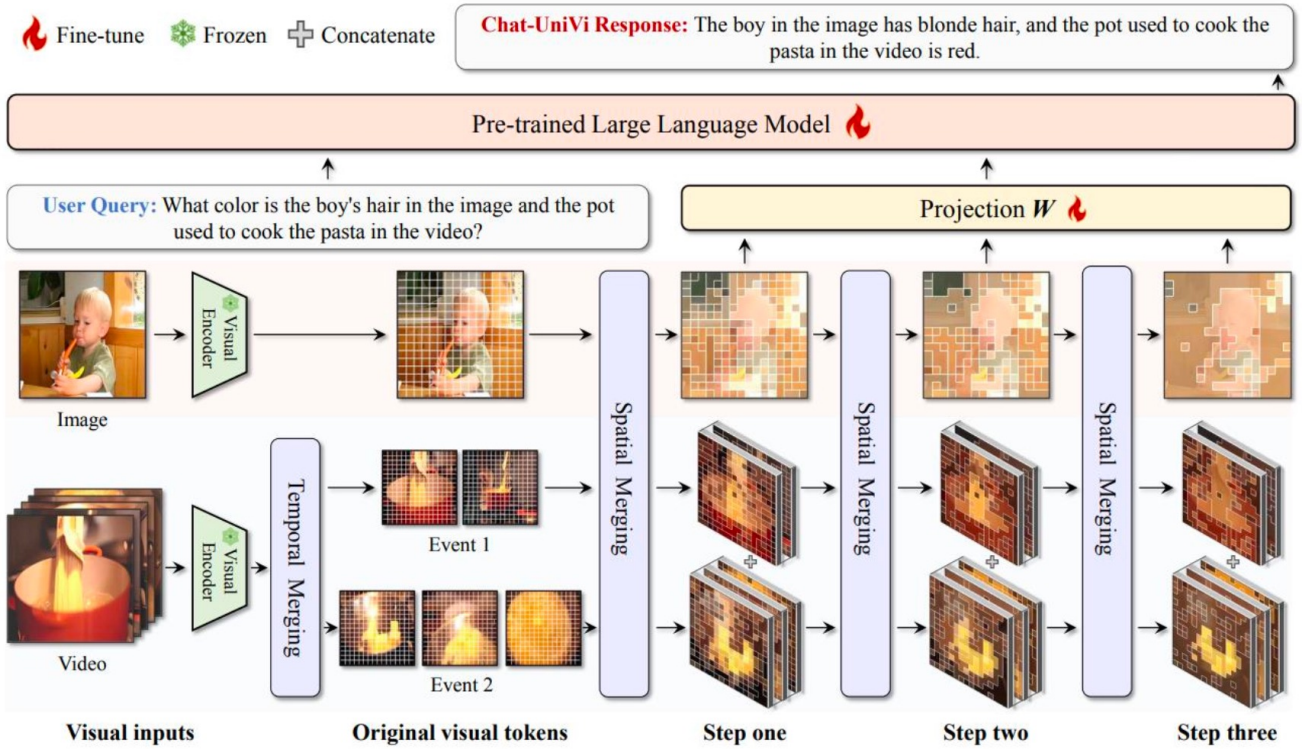

## EMU




[1] Generating Images with Multimodal Language Models. 2023
[2] Generative Pretraining in Multimodality. 2023

# MLLM Functionality: Text+Image+Video Generation

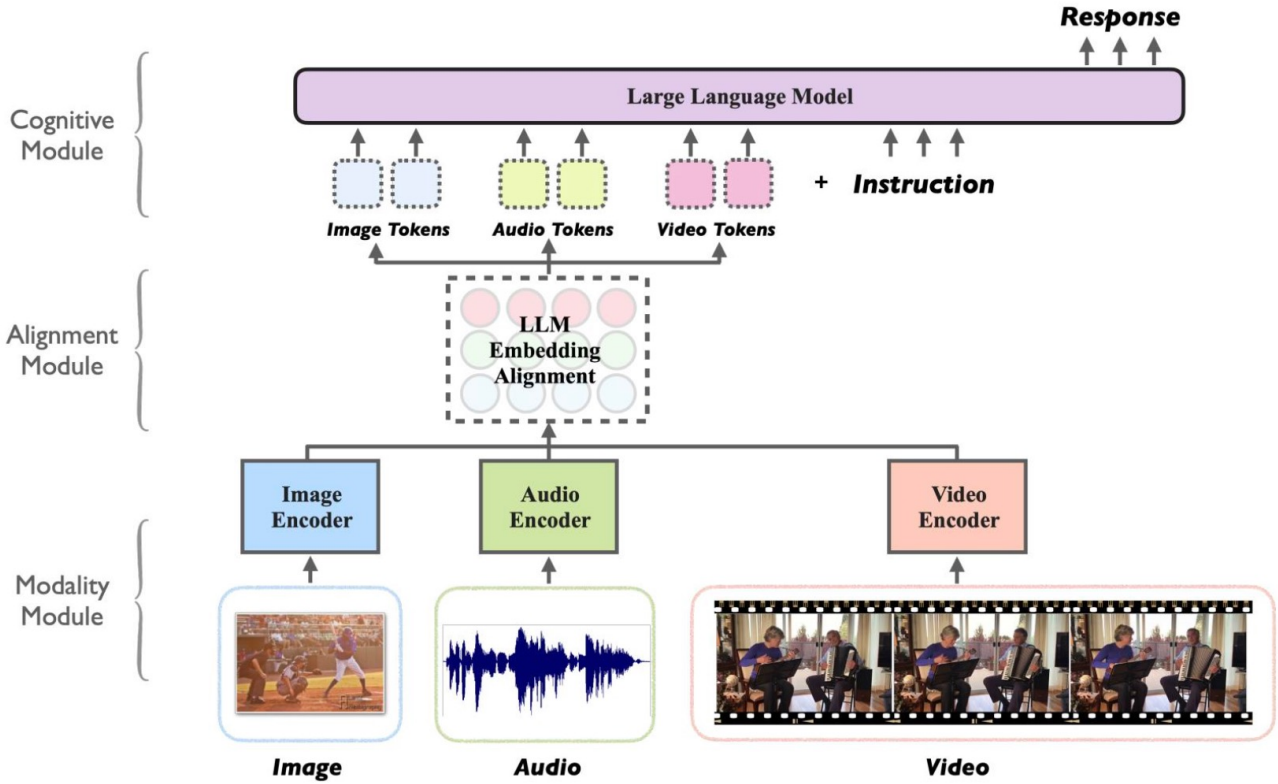
- Chat-UniVi




[1] Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. 2023

# MLLM Functionality: Text+Image+Video+Audio Generation

- Macaw-LLM



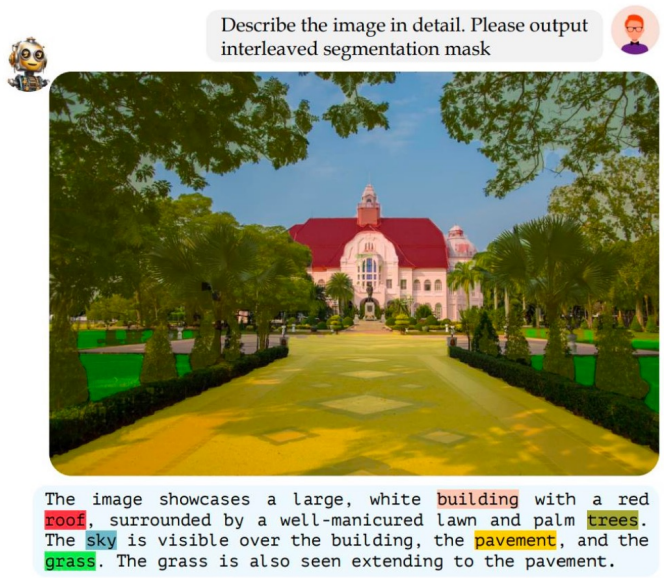

[1] Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. 2023

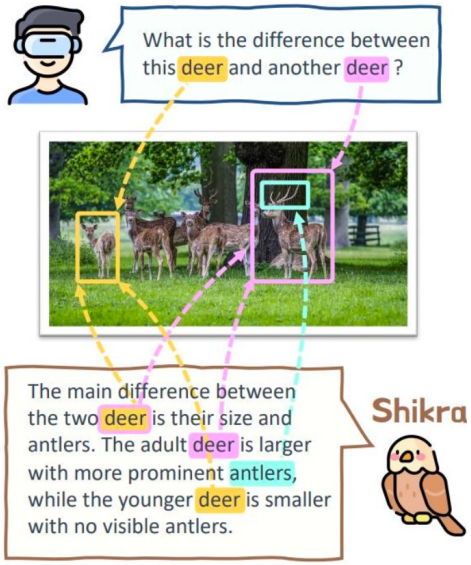
# MLLM Functionality: Fine-grained Capabilities

[Pixel-level Vision MLLM]

The vision MLLMs described above generally only support coarse-grained, instance-level visual understanding. This can lead to imprecise visual interpretations. Also due to the lack of visual grounding, these MLLMs will potentially produce hallucinations.

- Visual Grounding
- Visual Segmentation
- Visual Editing
- Visual Inpainting




Multimodal Large Language Models (MLLM) – Tutorial - https://mllm2024.github.io/CVPR2024/

# MLLM Functionality: Fine-grained Capabilities

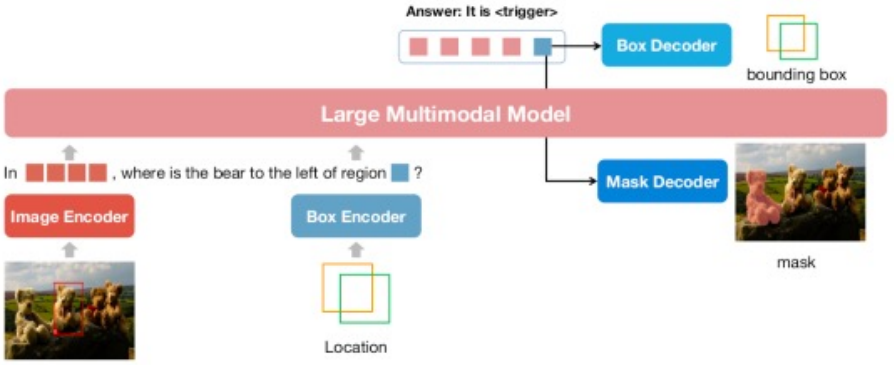
| | Modality (w/ Language) | | | |
|---|---|---|---|---|
| | Image | Video | Audio | 3D |
| Input-side Perceiving | Flamingo, Kosmos-1, Blip2, mPLUG-Owl, Mini-GPT4, LLaVA, InstructBLIP, VPGTrans, CogVLM, Monkey, Chameleon, Otter, Qwen-VL, GPT-4v, SPHINX, Yi-VL, Fuyu, … | VideoChat, Video-ChatGPT, Video-LLaMA, PandaGPT, MovieChat, Video-LLaVA, LLaMA-VID, Momentor, … | AudioGPT, SpeechGPT, VIOLA, AudioPaLM, SALMONN, MU-LLaMA, … | 3D-LLM, 3D-GPT, LL3DA, SpatialVLM, PointLLM, Point-Bind, … |
| | [Pixel-wise] GPT4RoI, LION, MiniGPT-v2, NExT-Chat, Kosmos-2, GLaMM, LISA, DetGPT, Osprey, PixelLM, … | [Pixel-wise] PG-Video-LLaVA, Merlin, MotionEpic, … | - | - |
| | Video-LLaVA, Chat-UniVi, LLaMA-VID | | - | - |
| | Panda-GPT, Video-LLaMA, AnyMAL, Macaw-LLM, Gemini, VideoPoet, ImageBind-LLM, LLMBind, LLaMA-Adapter, … | | | - |
| Perceiving + Generating | GILL, EMU, MiniGPT-5, DreamLLM, LLaVA-Plus, InternLM-XComposer2, SEED-LLaMA, LaVIT, Mini-Gemini, … | GPT4Video, Video-LaVIT, VideoPoet, … | AudioGPT, SpeechGPT, VIOLA, AudioPaLM, … | - |
| | [Pixel-wise] Vitron | | - | - |
| | NExT-GPT, Unified-IO 2, AnyGPT, CoDi-2, Modaverse, ViT-Lens, … | | | - |

Multimodal Large Language Models (MLLM) – Tutorial - https://mllm2024.github.io/CVPR2024/

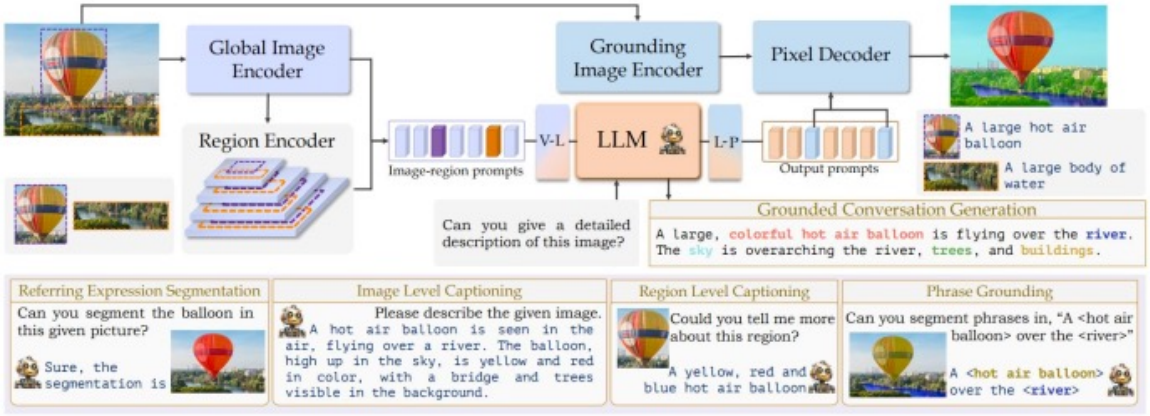
# MLLM Functionality: Image Fine-grained Capabilities

- Image-oriented Pixel-wise

  - NExT-Chat




  - GLaMM

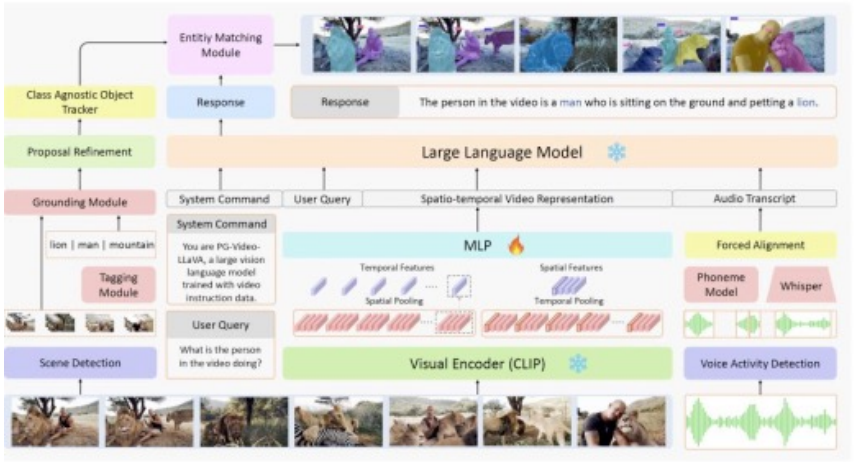



[1] NExT-Chat: An LMM for Chat, Detection and Segmentation. 2023
[2] GLaMM: Pixel Grounding Large Multimodal Model. 2023

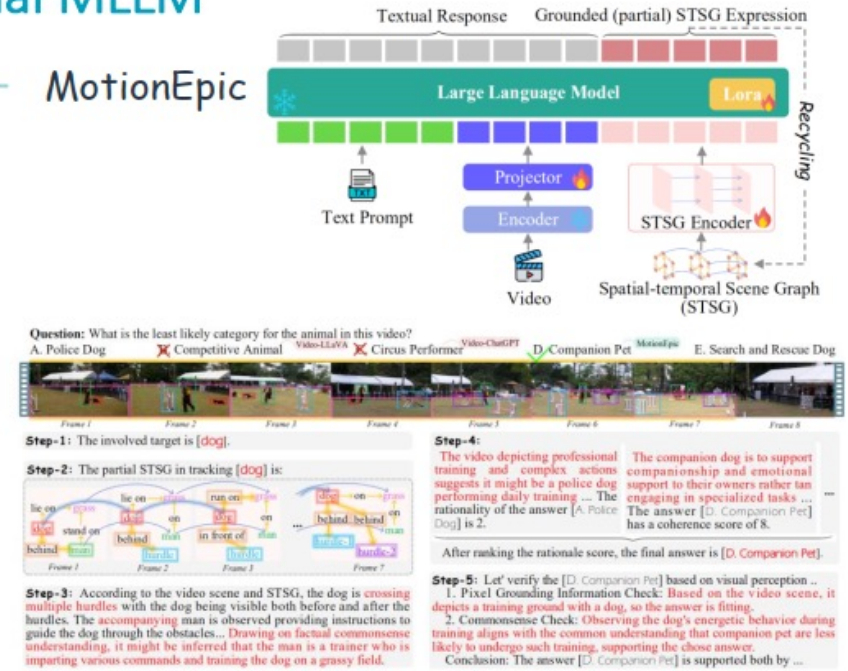
# MLLM Functionality: Video Fine-grained Capabilities

- Video-oriented Pixel-wise Regional MLLM

## PG-Video-LLaVA




## MotionEpic

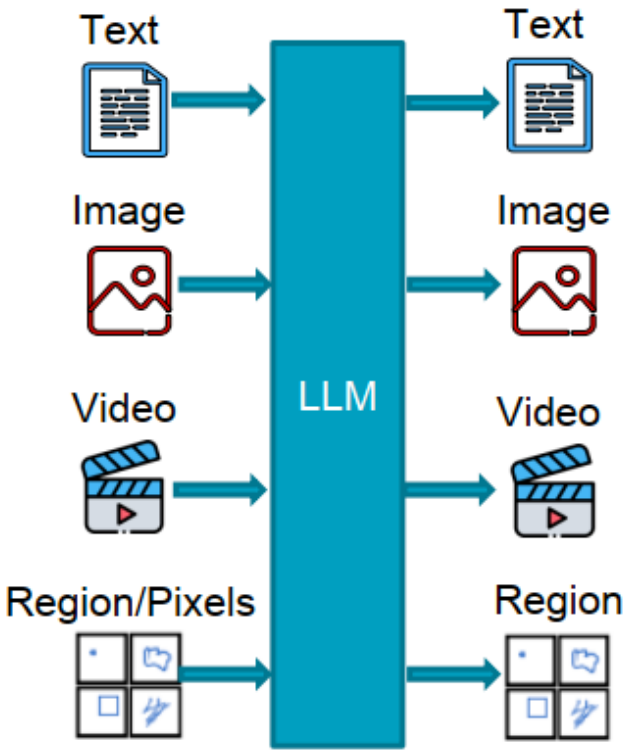



[1] PG-Video-LLaVA: Pixel Grounding in Large Multimodal Video Models. 2023
[2] Video-of-Thought: Step-by-Step Video Reasoning from Perception to Cognition. 2024

# MLLM Functionality: Unified Fine-grained Capabilities

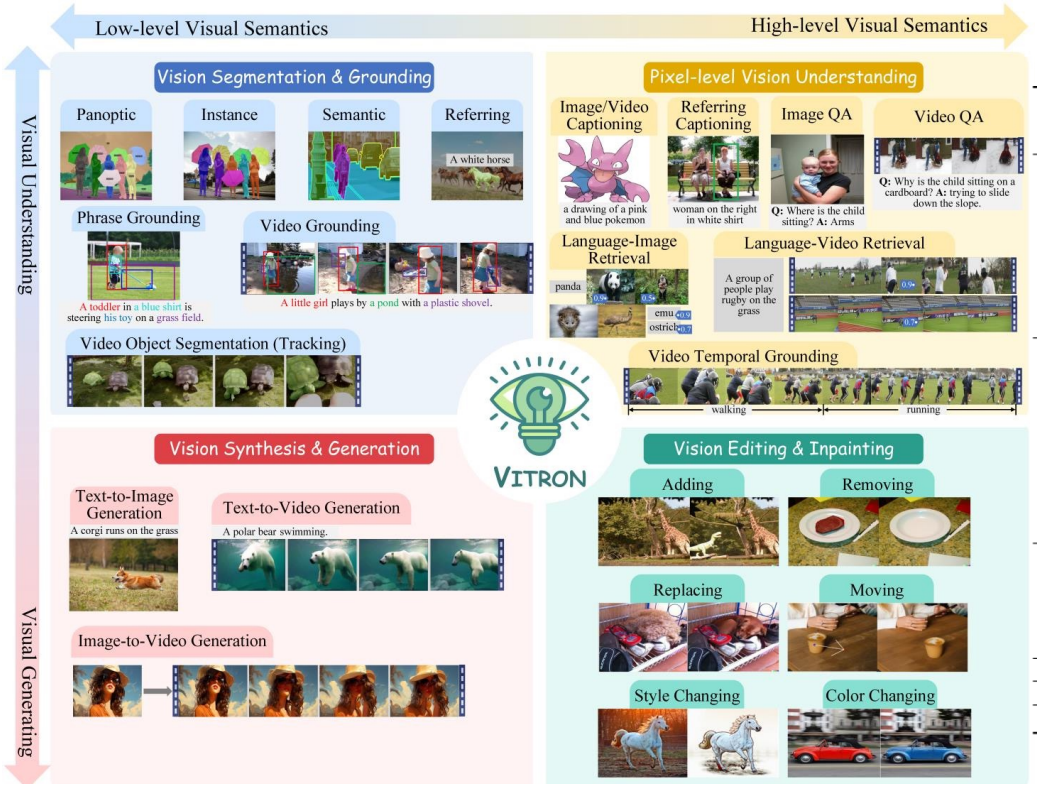
[Vitron]

Users input either an image or video (potentially specifying a region), and the LLM outputs content based on its understanding, generating, grounding or tracking the content to specific pixel-level regions of the image, video.




[1] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

# MLLM Functionality: Unified Fine-grained Capabilities

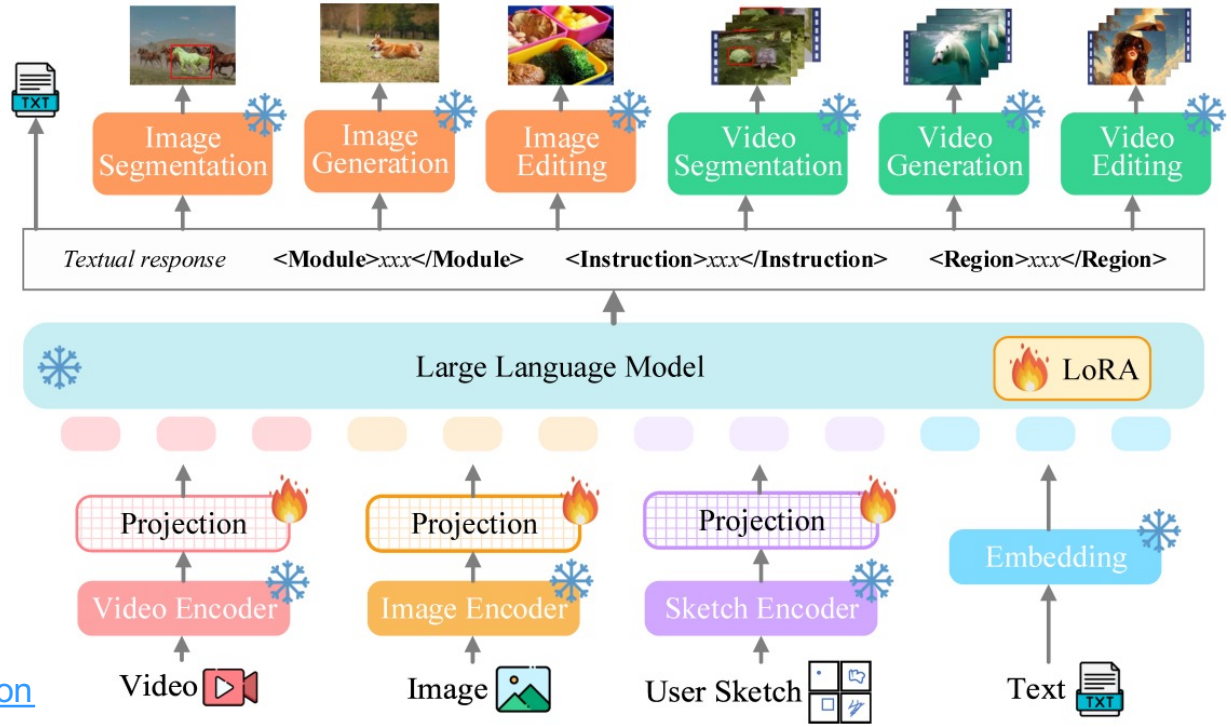



| Model | Vision Supporting | | Pixel/Regional Understanding | Segmenting/ Grounding | Generating | Editing |
|---|---|---|---|---|---|---|
| | Image | Video | | | | |
| Flamingo [1] | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| BLIP-2 [45] | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| MiniGPT-4 [126] | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| LLaVA [57] | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ |
| GILL [39] | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| Emu [90] | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| MiniGPT-5 [125] | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| DreamLLM [23] | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ |
| GPT4RoI [122] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| NExT-Chat [118] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| MiniGPT-v2 [13] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Shikra [14] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Kosmos-2 [72] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| GLaMM [78] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| Osprey [117] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| PixelLM [79] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ |
| LLaVA-Plus [58] | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ |
| VideoChat [46] | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Video-LLaMA [120] | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Video-LLaVA [52] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Video-ChatGPT [61] | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| GPT4Video [99] | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ |
| PG-Video-LLaVA [67] | ✗ | ✓ | ✓ | ✓ | ✗ | ✗ |
| NExT-GPT [104] | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ |
| VITRON (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

[1] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

# MLLM Functionality: Unified Fine-grained Capabilities

Vitron

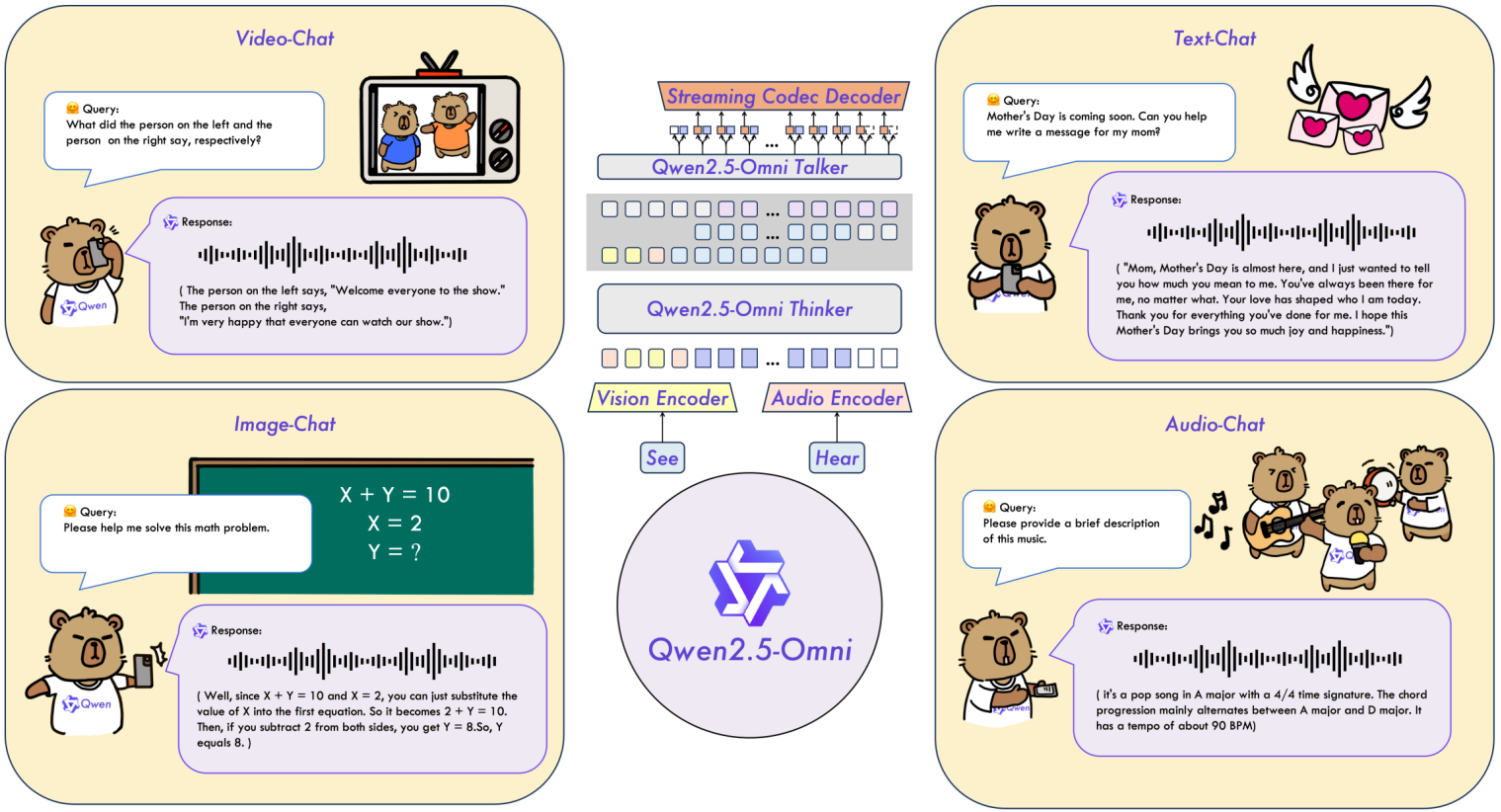



Project: https://vitron-llm.github.io/
Paper: https://is.gd/aGu0VV
Code&Demo: https://github.com/SkyworkAI/Vitron

[1] VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. 2024

# MLLM Functionality: Any-to-any VLMs

## Qwen2.5-Omni




https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://qwen.ai/blog?id=qwen2.5-omni

# MLLM Functionality: Any-to-any VLMs

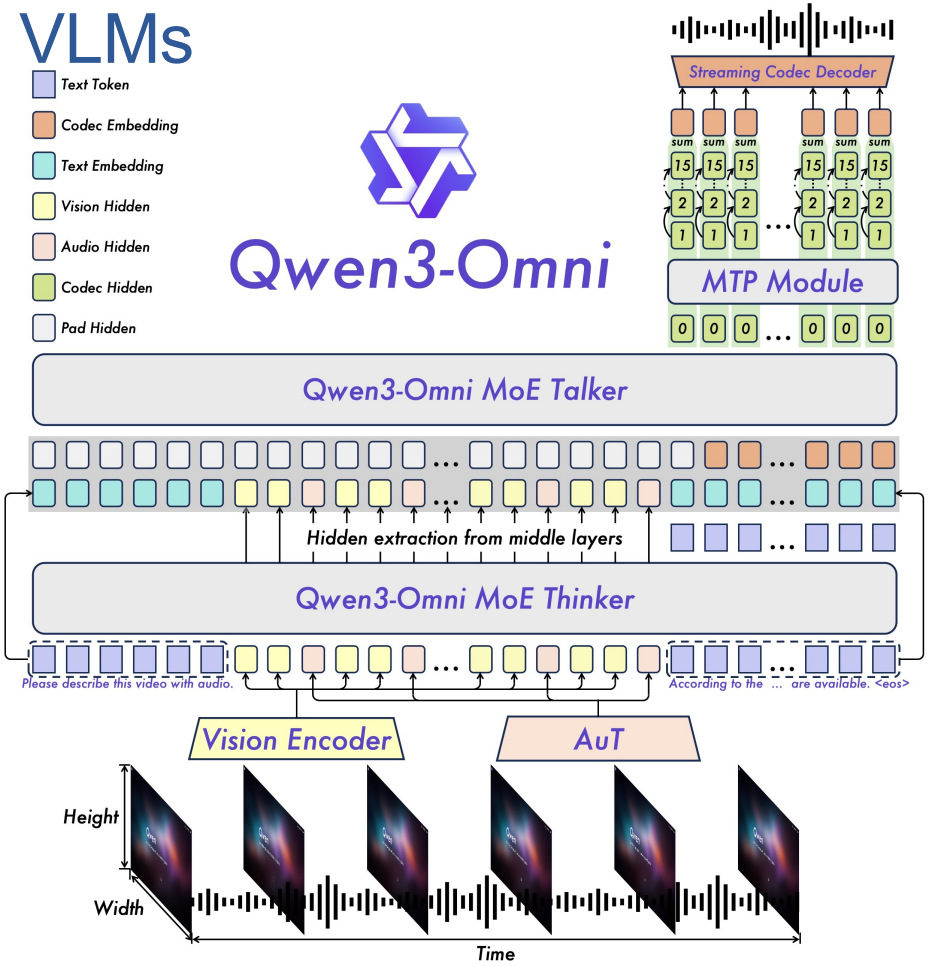
## Qwen3-Omni

Input Modalities:

- Text: Supports 119 languages.
- Image: Improved visual reasoning and OCR.
- Audio: Processes recordings over 40 minutes long; supports speech input in 19 languages.
- Video: Enhanced audio-visual synchronization (lip-syncing/scene reasoning).

Output Modalities:

- Text: Includes a “Thinking” mode for chain-of-thought reasoning.
- Audio (Speech): Higher-quality synthesis in 10 languages

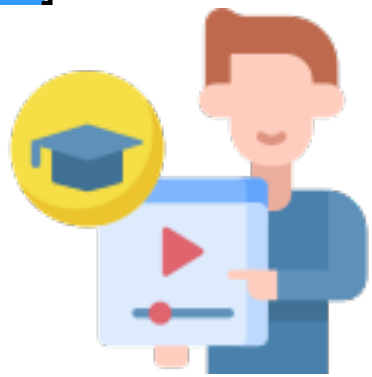



https://www.cs.cmu.edu/~mgormley/courses/10423/

Figure from https://qwen.ai/blog?id=65f766fc2dcba7905c1cb69cc4cab90e94126bf4

# Tutorials

- Introduction to Vision-Language Models in Python [link]
- Vision-Language Models [link]



# OMET Surveys



CS 1674 - INTRO TO COMPUTER VISION - 1010 - Lecture

https://go.blueja.io/dD9cADI1XEC7GIDZYwA3Dg

CS 2074 - INTRO TO COMPUTER VISION - 1010 - Lecture



https://go.blueja.io/BVvU9Nlfb0eubTeG5j8Hpw